Epidemic Forecasting on Networks: Bridging Local Samples with Global Outcomes

Yeganeh Alimohammadi[1], Christian Borgs[2], Remco van der Hofstad[3], and Amin Saberi[4]

[1] *Stanford University,* yeganeh@stanford.edu
[2] *University of California Berkeley,* borgs@berkeley.edu
[3] *Eindhoven University of Technology,* r.w.v.d.hofstad@tue.nl
[4] *Stanford University,* saberi@stanford.edu

Abstract

This paper studies Susceptible-Infected (SI), Susceptible-Infected-Removed (SIR), and related epidemic models in which infected individuals transition to an absorbing state, such as recovery or permanent infectiousness. In addition to infectious diseases, these models are used for studying the diffusion of innovations in which new behaviors, opinions, conventions, and technologies propagate from person to person through a social network.

We focus on the key challenge of forecasting epidemic trajectory and outbreak sizes and show that they can be predicted with a few samples from the network data. To this end, we propose a local algorithm for epidemic estimation, and prove the estimator's accuracy for both deterministic finite graphs and random networks, given certain neighborhood constraints. Further, leveraging the theory of local graph limits, we relate the time evolution in a sequence of graphs converging locally in probability with the epidemic in the limit graph. Finally, we validate our findings with experiments on synthetic models and real-world networks, such as Copenhagen and San Francisco's SafeGraph data.

1 Introduction

Epidemic models, originally conceived by Bernoulli (1760) and gaining prominence during the Spanish flu epidemics in the early 20th century (Ross and Hudson, 1917; Kermack and McKendrick, 1927), are developed for and frequently applied to the analysis of the spread of infectious diseases. These models categorize populations into various compartments such as Susceptible (S), Infectious (I), or Recovered (R), representing potential flow patterns individuals may experience throughout the course of an epidemic. As instrumental tools in public health policy formulation, they assist in forecasting various aspects of an epidemic, including its spread, the total number of infections, and its duration (Eubank et al., 2004; Larson, 2007; Lloyd-Smith et al., 2009; Heesterbeek et al., 2015; Scarpino and Petri, 2019). Furthermore, they aid in evaluating the potential impacts of health interventions, serving as a guide in optimizing strategies such as the allocation of limited vaccine resources or targeted closures, thus playing a central role in assessing the efficacy and selection of countermeasures during public health emergencies (Wu et al., 2005; Mamani et al., 2013; Kaplan, 2020; Bastani et al., 2021; Birge et al., 2022; Acemoglu et al., 2023).

Extending beyond the scope of infectious diseases, epidemic models play a pivotal role in studying the diffusion of innovations across social networks where new behaviors, opinions, conventions, and technologies spread from person to person (Bass, 1969). Understanding the dynamics of such adoptions within the underlying social networks can provide insights into potential "word-of-mouth" effects and the influence of decisions made by peers and colleagues. Originating from social science research, this method of analysis has informed our understanding of topics such as the diffusion of medical and agricultural innovations, the sway of viral marketing on new product success, cascading failures in power systems, and the dissemination of misinformation, presenting a critical avenue for anticipating the trajectory and impact of evolving trends and phenomena (Kalish and Lilien, 1983; Ford et al., 2006; Feder and Umali, 1993; Lee et al., 2015; Amini

and Minca, 2016; Yang et al., 2017; Mostagir and Siderius, 2023). Recent research further explores the fundamental algorithmic problems in these systems, aiming to utilize network data to optimize marketing strategies targeting influential network members, thereby enhancing the adoption rate of new products (Kempe et al., 2003; Goel et al., 2016; Lobel et al., 2017; Ajorlou et al., 2018; Akbarpour et al., 2018; Manshadi et al., 2020; Chin et al., 2022).

Predicting the trajectory of the epidemic is a central challenge for the above applications, primarily approached through modeling-based techniques and simulation methods. The common mean-field models, rooted in deterministic or stochastic PDEs, simplify the analysis by assuming uniform mixing among individuals, but often overlook the intricacies of infection transmission networks (Bartlett, 1949; Britton et al., 2019; Dimitrov and Meyers, 2010; Mukherjee and Seshadri, 2022). Random network models address this issue, with numerous models rigorously studied in the literature (Bampo et al., 2008; Manshadi et al., 2020; Kiss et al., 2017). However, selecting the right model for a specific population poses challenges, and parameter fitting can cause major prediction variations if misspecified. Alternatively, simulation methods utilize granular data like individual-level mobile tracking to trace epidemic progression (Bajardi et al., 2011; Wesolowski et al., 2012; Chang et al., 2021). However, acquiring complete data is difficult, with these approaches also presenting privacy concerns and lack of robustness when faced with incomplete data.

Here, we propose a distinct, data-driven approach that relies on the collection of small samples of network data to predict an epidemic. Recent work has ventured into similar methodologies, particularly in seeding strategies and estimating the final size of outbreak, (Eckles et al., 2022; Alimohammadi et al., 2022). However, the scope of their work is more constrained than what we present here. For an in-depth comparison, see Section 1.2.

Our proposed algorithm uses samples from the interaction networks, sufficient for approximating the future trajectory of an epidemic. It leverages the local neighborhood of randomly chosen individuals to create an estimator for the proportion of the population in each state of the epidemic at any given time. The analysis of the algorithm provides a bound on the estimator's error on any fixed graph. Importantly, under certain assumptions, we offer an upper bound on the sample size necessary for tracing the time evolution of the epidemic within an ε additive error; this bound is independent of the underlying network size. Our findings, applicable to both deterministic graphs and a broad class of random network models, suggest that accessing the local network structure of a few individuals is sufficient to estimate the time evolution of the epidemic.

1.1 Summary of Our Contribution

Our work makes several key contributions:

1) Algorithmic Estimation: We introduce a *local algorithm* for estimating the epidemic's time evolution in a network by simulating the epidemic process in the neighborhood of a selection of random nodes going backward in time. To describe the algorithm's dynamics, we use the Susceptible-Infectious-Recovered (SIR) model as an illustrative example. Under the SIR model, an infected node recovers after a time drawn from an arbitrary distribution and, while infected, transmits the disease to its neighbors at a fixed rate. Given an initial node v and an exploration budget k, the algorithm predicts v's status (susceptible, infectious, or recovered) at any time $t \geq 0$. The process starts from v and explores at most k nodes locally. For this purpose, the process determines node v's recovery time, and then draws recovery, and contact times for each adjacent node, allowing transmissions only if the respective contact time is earlier than the initiating node's recovery time. Then, the algorithm traverses transmission edges backward in a breadth-first manner until it reaches k nodes or exhausts all edges. Next, v's infection and recovery time can be computed using a directed distance metric derived from contact times, determining v's state over time for one instance of the process. By averaging the outcomes of this algorithm with independent starting points, we construct an estimator for the time evolution of epidemics. For details, see Section 2.1.

2) Theoretical Guarantees of the Estimator: We give exact bounds on the error of our estimators for any given deterministic graph (Theorem 2.1). Leveraging this result and under the assumption of moderate growth in local subgraph sizes (formalized by a tightness condition), we later prove that the required sample size for an ε additive error is a constant independent of the network size (Theorem 4.3). Second, we extend

our result to random network models, where we bound the error of our estimator under a condition ensuring that empirical distributions of local neighborhood structures are consistently retained across different network realizations (Theorem 2.3). Similar to our results on deterministic graphs, the sufficient sample size for the estimator depends solely on the desired prediction accuracy, suggesting that the time evolution of epidemics can be predicted by probing the local network structure of just a few individuals.

The latter result can be applied to various network models such as Erdös Rényi (Erdős et al., 1960), configuration model (Bollobás, 1980), preferential attachment model (Yule, 1925; Barabási and Albert, 1999), stochastic block model (Holland et al., 1983), and geometric random graphs (Gilbert, 1959). In the context of our findings, we demonstrate that the epidemic's time evolution concentrates around its mean, a property previously established for a specific class of random graph models (Janson et al., 2014; Decreusefond et al., 2012).

3) Asymptotic Characterization: We also study the asymptotics of epidemics on a sequence of graphs of growing sizes. Leveraging the theory of local graph limits (Benjamini and Schramm, 2001; Aldous and Steele, 2004), we prove that the epidemic's time evolution in a sequence of graphs with a local limit in probability converges to the same process in the graph limit (Theorem 2.5). This implies that the time evolution of epidemics is essentially a 'local' property of the graph.

4) Applicability to General Epidemic Models: Building on SIR models, our framework can be extended to accommodate more complex epidemic models that include various intermediate states, particularly those where an individual's level of infectiousness changes over time as they transition through different phases. Initially, a susceptible person might be infected but not contagious (i.e., exposed), then move through stages with fluctuating transmission rates, potentially leading to a recovery state. Further, we can consider settings in which the initial configuration is not uniform but is shaped by a probability measure influenced by characteristics of their local neighborhood, such as degree. Our model also has the capacity to incorporate interventions like vaccination or isolation strategies (see Section 2.3).

In the context of real-world applications, our framework encompasses several well-established epidemic models, including those with heterogeneous infectiousness for HIV (May and Anderson, 1987; Isham, 1988), malaria (Mandal et al., 2011; Gupta et al., 1994), models with carrier states for tuberculosis (Aparicio et al., 2000; Blower et al., 1995) and typhoid (Cvjetanović et al., 1971), influenza (Andreasen et al., 1997), COVID-19 (Bertsimas et al., 2021; Mukherjee and Seshadri, 2022), and even models of information cascade (Watts, 2002) and viral marketing models (Bass, 1969; Jackson and Yariv, 2005; Banerjee et al., 2013; Bampo et al., 2008; Ajorlou et al., 2018; Manshadi et al., 2020). A central theme behind these models is that they all reach an eventual absorbing state, whether infectiousness or recovery. So it is possible to run the epidemic process backward and apply our algorithm. The flexibility to capture this breadth of epidemiological models and dynamics makes our framework broadly applicable.

5) Experimental Validation: We empirically validate our theoretical finding by conducting experiments on synthetic and real-world networks, including the Copenhagen Interaction Network created by the Bluetooth data of over 400 students (Sapiezynski et al., 2019) and San Francisco SafeGraph data with more than 30,000 nodes representing census blocks and points of interests across San Francisco (Chang et al., 2021). Notably, the San Francisco SafeGraph dataset includes edge weights, representing the transmission strength between nodes. In our experiments, we compare the outcomes of running an SIR epidemic on the entire graph against our estimator's results, which uses local network information of a few nodes. Remarkably, even with access to less than 1% of the nodes in the San Francisco dataset and approximately 14% of nodes in the Copenhagen dataset, our methodology yields predictions that align closely with the actual epidemic trajectories, falling within the 95% confidence interval of the true time evolution (see details in Section 5).

1.2 Related Work

In the evolving landscape of epidemic prediction within operations research, a recent survey by (Gupta et al., 2022) highlights the paramount need for refined modeling approaches and efficient data collection mechanisms. Our research addresses this gap, focusing on the use of small samples of network data for more accurate

epidemic predictions. To position our contributions, we delve into two main avenues of related research: understanding epidemics with small data, and the asymptotic behavior of epidemics.

Prediction with small data: Harnessing the power of small data in epidemic modeling has gained attention in the past few years. Recent work by Baek et al. (2021) studied sample complexity for estimating diffusion models, including SIR, with limited data and unobservable networks. They derived lower bounds on the number of required samples to estimate outbreak size. Notably, without network information, their lower bounds increase with population size. This highlights the value of leveraging additional data layers, particularly network structure as in our method, to reduce sample requirements.

Recognizing the pressing need for network data, several studies showcase its pivotal role in enhancing predictions, especially regarding influence maximization. These studies typically lean on heuristics for probing the network data (Mihara et al., 2015; Stein et al., 2017; Chen et al., 2022) or offer theoretical perspectives tailored to specific random graph models (Wilder et al., 2018).

In this avenue, there are two closely related studies (Eckles et al., 2022; Alimohammadi et al., 2022). Both investigate the use of minimal network data under the independent cascade model, a variant of the SIR model where nodes transmit disease or information at a fixed probability p. The primary goal of (Alimohammadi et al., 2022) is to approximate the epidemic's final size, closely related to the size of the largest component under percolation — a mathematical structure where each edge of the graph is retained with a given probability p. They propose a local algorithm for this purpose. Their algorithm, bearing some resemblance to ours without a temporal element, starts from a random node and outputs if it can infect a constant number of other nodes. Under the assumption that the graph is an expander, they prove that constant queries to this algorithm yield an $(1-\epsilon)$ approximation of the infection's end size, drawing insights from their earlier research (Alimohammadi et al., 2023) on expanders and percolation.

Contrasting with their work, our focus extends beyond the final infection size to include the epidemic's time evolution. Additionally, our scope encompasses diverse epidemic models, whereas theirs remains restricted to SIR with constant recovery time. Furthermore, their result is limited to the more specific class of expander graphs. This restriction stems from the fact that their model initiates epidemics from a single node. In contrast, our approach assumes the epidemic begins from a small, fixed fraction of the entire population.

Turning our attention to the insightful work of Eckles et al. (2022), they focus on identifying optimal seeds under the independent cascade model with a fixed seeding budget. They offer methods to approximate the optimal seeding strategy while obtaining as minimal network information as possible. Their strategy is twofold: First, using an oracle that reveals an infection's ultimate spread from a chosen node, they show that limited queries can achieve an error of ε relative to the optimal seeding solution. In the second scenario, where observing edges is costly, they propose a probing algorithm that finds optimal seeds by querying a constant fraction of the total edges. The common thread between our work and theirs is the aim to use minimal network information for diffusion tasks. However, the specific goal of optimizing seeds is not within the scope of our study.

Another different avenue in the relation of the epidemic on networks and limited data has focused on inferring the global network structure from the limited observations of epidemic trajectories. For example, Graham (2008); Goldsmith-Pinkham and Imbens (2013); Netrapalli and Sanghavi (2012) set out to reconstruct networks, fitting parameters with observed data and operating on the assumption that the intrinsic network draws from a stochastic block model (or what they call a linear-in-mean model). Further expanding on this line of inquiry, Kim et al. (2014); Drakopoulos and Zheng (2017) adapted the approach to make use of dynamic epidemic data. In our work, rather than learning network model parameters from existing data, we are interested in data collection to make predictions about epidemics.

Asymptotic Analysis of Epidemics: Many studies have built rigorous foundation on concentration of time evolution of epidemics under different random graph models, including Erdös-Rényi graphs (Budhiraja et al., 2012; Coppini et al., 2020), configuration models (Janson et al., 2014; Decreusefond et al., 2012), and their dynamic variants (Jacobsen et al., 2016; Ball and Britton, 2022). Additionally, insights on the duration of epidemics have been illuminated by works like (Bhamidi et al., 2014; Lashari et al., 2021). For a more comprehensive overview of mathematical models of epidemics, the book by (Kiss et al., 2017) serves as a great resource. Many of these random network models meet the tightness and stable neighborhood conditions required by our theorem, thus our concentration results in Theorem 2.3 are applicable to them.

Most recently, the beautiful works of (Lacker et al., 2019; Ganguly and Ramanan, 2022) study a general class of processes on locally convergent graphs, showing that given certain conditions, the empirical node state distribution converges to the limit. The focus of (Lacker et al., 2019) is on diffusion processes, whereas (Ganguly and Ramanan, 2022) brings jump processes into the fold — a class of diffusion processes that can fit the SIR model. While they study a more general random process on graphs, the applicability of their results is limited to a narrower set of graph structure, such as those with bounded maximum degrees or those converging locally to a generalized branching process. In contrast, our convergence theorem (Theorem 2.5) does not necessitate assumptions about bounds on the maximum degree or a specific structure in the limit.

2 Model and Main Results

To facilitate a clear presentation, we first present our main results for a classic Susceptible-Infected-Removed (SIR) model where individuals can be susceptible, infected, or recovered. In this model, individuals are represented as nodes in a graph, with edges representing potential transmissions. An infected node recovers at a time drawn from an arbitrary recovery time distribution D_R, and while infected, it transmits the disease to its neighbors following a Poisson time process with a fixed rate (denoted by D_I), after which the infected node recovers. We consider a scenario where each node is independently infected initially with a probability greater than zero, denoted as $\rho > 0$.

Our results extend beyond the classic SIR model. We prove our results on a general model of epidemics with time-varying infectiousness and heterogeneous initial conditions in Section 2.3.

2.1 Local Estimator for SIR Model

We introduce a local algorithm that uses the information of a small number of individuals in the local neighborhood of the node v to determine its state with respect to the epidemic at any time in the future $t \geq 0$. This is achieved by a backward simulation of the epidemic process.

Given a parameter k and an initial node v as input, the algorithm simulates the backward epidemic process starting from v until it reaches at most k other nodes in the neighborhood of v. The output is a vector $(S_{k,v}(t), I_{k,v}(t), R_{k,v}(t))_{t\geq 0}$, where $S_{k,v}(t)$, $I_{k,v}(t)$, and $R_{k,v}(t)$ are indicators showing whether node v is susceptible, infectious or recovered at time t under the simulated process.

For a more precise illustration of the backward process, we begin by determining the recovery time for node v. Subsequently, for each node u connected to v, we draw its recovery times r_u from D_R, and the contact time $c_{(u,v)}$ between u and v from D_I. A transmission along a directed edge (u, w) is feasible only when the contact time is less than the recovery time of the starting endpoint of the edge, i.e., $c_{(u,w)} < r_u$; otherwise, the edge gets discarded. The algorithm then performs a breadth-first search from node v, traversing edges backward until it either reaches k nodes or runs out of edges. During this traversal, the algorithm consistently computes recovery and transmission times for newly encountered nodes and edges.

Using the contact times as edge weight – and assigning ∞ where contact time exceeds recovery time– we define a directed distance $\text{dist}_{(T)}(x, y)$ between two nodes x, y as the length of the shortest weighted path from x to y. This distance conceptually represents the time it takes fornode x to infect y. Leveraging this construction, the algorithm determines v's infection time by finding the shortest path with respect to $\text{dist}_{(T)}$ from the observed initially infected nodes to v. Here, we assumed the knowledge of each observed node's initial state. However, without this knowledge, the initial state is determined by infecting each node with probability ρ. Furthermore, the recovery time of node v is obtained by adding r_v to the infection time of v. Finally, the algorithm outputs the state of node v across time. See the details in Algorithm 1 below.

We choose q independent uniform starting nodes $v_1, \ldots, v_q$, and apply Algorithm 1 to them to estimate the time evolution of the epidemic from these nodes. Assuming that initially each node is infected with probability ρ, we define the estimator

$$\hat{S}^{(\rho)}_{q,k,n}(t) = \frac{\sum_{i=1}^{q} S_{k,v_i}(t)}{q} \tag{1}$$

as the fraction of susceptible nodes in the q starting nodes at time t. Similarly, define $\hat{I}^{(\rho)}_{q,k,n}(t)$, $\hat{R}^{(\rho)}_{q,k,n}(t)$, as the fractions of infectious and recovered nodes at time t.

Algorithm 1: Local algorithm – the backward epidemic process

Input: Integer $k > 0$, root node v, initial infected nodes I_0, and G_n.
Let $O = \{\}, W = \{v\}$ be the list of observed and waiting to be observed nodes, respectively.
while $|O| \leq k$ *and* $W \neq \emptyset$ **do**
  Let u be the next node in W.
  for $w \in N(u)$ **do**
    Draw contact times for the edge $c_{(u,w)}$.
    if $r_w = \emptyset$ **then**
      Draw recovery time r_w from D_R.
    end
    if $r_w > c_{(u,w)}$ *and* $w \notin O$ **then**
      Add w to W.
    end
  end
  Remove u from W and add to O.
end
$\inf_v = \mathrm{dist}_{(T)}(O \cap I_0, v)$ (shortest path from $O \cap I_0$ to v using transmission edges).
$\mathrm{rec}_v = \inf_v + r_v$.
$S_{k,v}(t) = \mathbb{1}\{t \leq \inf_v\}$, $I_{k,v}(t) = \mathbb{1}\{\inf_v < t \leq \mathrm{rec}_v\}$, $R_{k,v}(t) = \mathbb{1}\{t > \mathrm{rec}_v\}$.
Output: $(S_{k,v}(t), I_{k,v}(t), R_{k,v}(t))_{t \geq 0}$

2.2 Efficacy of the Local Estimator

In the following sections, we rigorously examine the error of our local estimators, $\hat{S}^{(\rho)}_{q,k,n}(t)$, $\hat{I}^{(\rho)}_{q,k,n}(t)$, and $\hat{R}^{(\rho)}_{q,k,n}(t)$ in predicting the time evolution of epidemics $S^{(\rho)}_n(t)$, $I^{(\rho)}_n(t)$, and $R^{(\rho)}_n(t)$. We start by presenting an exact bound for the accuracy of the estimator with a predetermined number of queries q and input size k in finite deterministic graphs (see Section 2.2.1). In Section 2.2.2, we extend these insights to random network models, and in Section 2.2.3 we extend our findings to sequences of growing graphs with similar local structures, leveraging the theory of local graph limits.

2.2.1 Finite Deterministic Graph.

Given a fixed deterministic graph G_n, we define the vector $\mathscr{E}^{(\rho)}_n(t) = (S^{(\rho)}_n(t), I^{(\rho)}_n(t), R^{(\rho)}_n(t))$ representing the epidemic state at time t. Each of the components $S^{(\rho)}_n(t)$, $I^{(\rho)}_n(t)$ and $R^{(\rho)}_n(t)$ are random variables representing the proportion of susceptible, infectious, and recovered nodes in G_n, respectively. Similarly, define $\hat{\mathscr{E}}^{(\rho)}_{q,k,n}(t) = (\hat{S}^{(\rho)}_{q,k,n}(t), \hat{I}^{(\rho)}_{q,k,n}(t), \hat{R}^{(\rho)}_{q,k,n}(t))$ as the estimator vector, obtained from running Algorithm 1 with input k and q independent starting points. We can directly bound the error of the estimator using the following expression:

$$\varepsilon_r(G_n, k) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{|B_r(G_n, v)| > k\},$$

where $B_r(G_n, v)$ is the subgraph of G_n containing all nodes at a graph distance of at most r from v. This expression captures a *tightness* condition on the neighborhood sizes around uniform random nodes.

Theorem 2.1 (Local Estimator for a Finite Graph)**.** *Let G_n be a deterministic graph of size n. Consider an SIR epidemic in which each node is initially infected with an independent probability of $\rho > 0$. Then for any $t \in [0, \infty]$,*

$$\mathbb{P}\Big(|\mathscr{E}^{(\rho)}_n(t) - \mathbb{E}[\mathscr{E}^{(\rho)}_n(t)]| > \delta\Big) \leq \frac{16}{n\delta^2} + \min_{r,k \geq 1} \Big(\frac{k}{n} + \frac{16\varepsilon_{2r}(G_n, k)}{\delta^2} + \frac{(1-\rho)^r}{\delta}\Big). \tag{2}$$

Further, the error of the estimator is bounded, i.e., for any $t \in [0, \infty]$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{n,q,k}(t) - \mathscr{E}^{(\rho)}_n(t)| > \delta\Big) \leq \frac{32(k+1)}{\delta^2 n} + 2e^{-2q\delta^2} + \frac{2}{\delta} \min_{r \geq 0} \big((1-\rho)^r + (1 + \frac{16}{\delta})\varepsilon_{2r}(G_n, k)\big). \tag{3}$$

In (2) and (3), the probability is over both the randomness of the algorithm and the epidemic process.

Our theorem presents an upper bound on the estimator's error. This bound consists of multiple terms, reflecting the nuanced interplay between various parameters. A main component of our bound depends on the interrelation between the radius r and and the fraction of nodes with r-neighborhood larger than k, expressed as $\varepsilon_r(G_n, k)$. We can bound this term in many scenarios. For example, consider the case that G_n has maximum degree Δ. Then the r neighborhood of the node has at most Δ^r nodes, so we can choose $r < \log_\Delta(k)$ small enough to deduce $\varepsilon_{\lfloor \log_\Delta(k) \rfloor}(G_n, k) = 0$. As a result, an upper bound for the error in (3) simplifies to a function of number of queries and input k, as $\frac{32(k+1)}{\delta^2 n} + 2e^{-2q\delta^2} + \frac{2}{\delta}(1-\rho)^{\log_\Delta(k)}$.

What happens if the graph does not have a maximum degree limit, meaning that the maximum degree in G_n may grow quickly with n? Our theorem can still imply that with small k and q the error is small, provided that the size of the local neighborhood of a uniform random node grows slowly. One can ensure such behavior by implementing a constraint on bounding $\varepsilon_r(G_n, k)$ for large enough constant k. We refer to this specific constraint as tightness, elaborated in Definition 3.1. In Theorem 4.3, we prove that under the tightness condition, and for any precision δ, there are constant q and k such that the estimator achieves δ-additive error with probability at least $1 - \delta$.

Remark 2.2 (Local estimator of the final size of the epidemic)**.** Since the guarantees of the theorem are independent of t, they also give guarantees for the final size of the infection. ◀

2.2.2 Finite Random Graphs

To determine the error margin of the estimator for random graphs, we begin by illustrating how our algorithm accommodates random network models. Two approaches can be adopted without altering our results. One option is to first draw a network realization from the desired model and then feed local samples to the algorithm. Alternatively, the network can unfold locally on-the-fly around the input node during the backward process rather than fixing the full network realization upfront. Both strategies yield the results we aim to showcase.

To prove the accuracy of the estimator, we require a condition called *Stable Neighborhood Structure* (see Definition 3.2). This condition ensures that distributions of local network structures are similar across random instances. For example, a model that randomly generates either a complete graph or isolated nodes with equal probability would not satisfy this condition, since the local structures differ drastically between those cases. With this condition met, we demonstrate that for a specified precision $\delta > 0$, there exist constants k_δ and q_δ with respect to the network size to ensure an estimation error is at most δ:

Theorem 2.3 (Local Estimators for Random Graphs)**.** *Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of random graphs satisfying tightness and stable neighborhood structures (Definitions 3.1 and 3.2). Then $\mathscr{E}_n^{(\rho)}(t)$ is concentrated,*

$$\left|\mathscr{E}_n^{(\rho)}(t) - \mathbb{E}[\mathscr{E}_n^{(\rho)}(t)]\right| \xrightarrow{\mathbb{P}} 0 \quad \textit{as} \quad n \to \infty.$$

Furthermore, given any $\delta > 0$, there exists constants N, q, k such that for any $n > N$, and any $t \geq 0$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}_{n,q,k}^{(\rho)}(t) - \mathscr{E}_n^{(\rho)}(t)| > \delta\Big) \leq \delta. \tag{4}$$

Policy implications of this result arise in scenarios where social planners might lack access to the specifics of the social interactions (even for the few local samples required by Theorem 2.1). Alternatively, the planner may wish to model the interaction networks, perhaps under varying intervention policies, and implement a rapid algorithm to predict how an epidemic might unfold. Our result shows that Algorithm 1 can be implemented fast, since it needs to be run on a few nodes. This would enable planners to test and compare different policies efficiently.

Remark 2.4 (Application to Random Graph Models)**.** The stable local neighborhood and tightness assumptions apply to most sparse random network models. In particular, we will show that both conditions hold for graphs converging locally in probability (see Appendix C.3). Using this insight, we can apply our results to configuration models (Dembo and Montanari, 2010), sparse inhomogeneous random graphs (including stochastic block models) (Bollobás et al., 2007), preferential attachment models (Berger et al., 2014; Garavaglia et al., 2022), random intersection graphs (Kurauskas, 2022; van der Hofstad et al., 2021),

random graph models with communities (Trapman, 2007; Ball et al., 2010; van der Hofstad et al., 2015), and spatial inhomogeneous random graphs (van der Hofstad et al., 2023). The latter includes hyperbolic random graphs (Krioukov et al., 2010; Komjáthy and Lodewijks, 2020). For an in-depth exploration of various network models and their corresponding limits, we recommend the book by van der Hofstad (2024). ◀

2.2.3 Sequence of Growing Graphs

We generalize our previous results using the theory of local graph limits (Benjamini and Schramm, 2001; Aldous and Steele, 2004). Intuitively, a sequence of (possibly random) graphs $G_n\}_{n\in\mathbb{N}}$ is said to have a *local limit in probability* if the empirical distributions of the neighborhoods of randomly sampled nodes converge in probability. The limit is then a probability measure μ on the space $\mathscr{G}_\star$ of rooted, locally finite graphs. We will use (G, o) for a graph G with root o in $\mathscr{G}_\star$. See Section 3.3 for the precise definitions. We prove that epidemics exhibit well-defined limit behavior under local convergence. Further, the final size and the time evolution of an epidemic in finite graphs converge to those on the limit graph.

Theorem 2.5 (Convergence of the Epidemic Processes)**.** *Let $(G_n)_{n\geq 1}$ be a graph sequence that converges locally in probability to $(G, o) \sim \mu$, where μ is a deterministic probability measure over $\mathscr{G}_\star$. Then, there are functions $\mathscr{E}(t) = (s(t), i(t), r(t))$ such that, for any $t \geq 0$ $\mathscr{E}_n(t) \xrightarrow{\mathbb{P}} \mathscr{E}(t)$, and further, $R_n^{(\rho)}(\infty)/n \xrightarrow{\mathbb{P}} r(\infty)$.*

One implication of this result is that epidemics are essentially a *local property* of the graphs. As a result, it is possible to relate epidemic dynamics across networks with shared local structures but vastly differing scales.

In Theorem 2.5, the functions $s(t)$, $i(t)$ and $r(t)$ can be expressed in terms of the limit graph:

$$s(t) = \mu(o \in \mathcal{S}^{(\rho)}(t)), \qquad i(t) = \mu(o \in \mathcal{I}^{(\rho)}(t)), \qquad r(t) = \mu(o \in \mathcal{R}^{(\rho)}(t)),$$

where $(\mathcal{S}^{(\rho)}(t), \mathcal{I}^{(\rho)}(t), \mathcal{R}^{(\rho)}(t))$ are the sets of susceptible, infected and recovered nodes for an epidemic on (G, o) started from $\mathcal{R}^{(\rho)}(0) = \varnothing$, and every node is in $\mathcal{I}^{(\rho)}(0)$ independently with probability ρ. Here, the final size of the epidemic can be described by taking the time to infinity, $R_n^{(\rho)}(\infty) = \lim_{t\to\infty} R_n^{(\rho)}(t)$. Equivalently, the final size of the epidemic can also be described as $\mu(o \in \mathcal{R}^{(\rho)}(\infty)) = \mu(\mathscr{C}^-(o) \cap \mathcal{I}^{(\rho)}(0) \neq \varnothing)$, where $\mathscr{C}^-(o)$ is the set of all nodes reached by the backward epidemic process started from o. As part of the proof of the theorem, we show that the epidemic functions $(s(t), i(t), r(t))$ are well-defined on the limit graph.

Algorithmic insights in the limit: In our proofs, we will show that any sequence of locally convergent graphs satisfies the tightness and stable neighborhood conditions (Definitions 3.1 and 3.2). As a consequence, we can put the three theorems together to yield the local estimate of the epidemic on the limit graph $\mathscr{E}(t) = (s(t), i(t), r(t))$.

Theorem 2.6 (Local Estimation of the Limit)**.** *Assume $(G_n)_{n\in\mathbb{N}}$ and the epidemic process satisfy the conditions of Theorem 2.5. Then given $\delta > 0$, there exists constants k q such that for any $t \geq 0$*

$$\limsup_{n\to\infty} \mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{q,k,n}(t) - \mathscr{E}(t)| > \delta\Big) \leq \delta. \tag{5}$$

Similarly, at time $t = \infty$,

$$\limsup_{n\to\infty} \mathbb{P}\Big(|\hat{R}^{(\rho)}_{q,k,n}(\infty) - r^{(\rho)}(\infty)| > \delta\Big) \leq \delta. \tag{6}$$

2.3 Generalization to Other Models

The core results presented in this work, specifically the convergence of epidemic processes in Theorem 2.5, generalize in several important ways. In this section, we explore the applicability of our results to various epidemic models and starting configurations, providing a more comprehensive picture of the theorem's reach.

2.3.1 General Epidemics

We introduce a generalized epidemic model with time-varying infectiousness that would apply to SI, SEIR, or different variations of it, as well as, to intervention strategies such as vaccination and isolation strategies.

Time-varying infectiousness: We explore a model where a node's infectiousness varies over time, accommodating stages like exposure or fluctuating infectiousness levels (see Figure 1). This model distinguishes two timescales: 1) the *epidemic timescale* tracking disease progression network-wide, and 2) *node-specific timescales* beginning at each node's contraction from neighbors.

In this model, nodes are either susceptible or occupy a disease state from $\mathscr{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$. This set describes the m sequential states a node v undergoes after contracting the disease, starting with $\mathcal{D}_1$ and subsequently moving to $\mathcal{D}_2$, and so forth. Once a node in the disease state transmits the disease to its susceptible neighbor, that neighbor enters the $\mathcal{D}_1$ state, marking the beginning of its node-specific timescale.

The progression between disease states partitions the node-specific timescale $[0, \infty]$ into m intervals $[t_0, t_1), [t_1, t_2), \cdots [t_{m-1}, \infty]$, where the interval $[t_{i-1}, t_i)$ corresponds to the state $\mathcal{D}_i$. The transition times between disease states, $t_1 \leq t_2 \leq \ldots \leq t_m$, are drawn from a distribution $\tau_v : \{t_1, t_2, \ldots, t_m\} \to \mathbb{R}_+^m$. For example, Figure 1 shows a scenario where disease states are Exposed, Infectious, Quarantine, and Recovered. The node-specific timescale is divided into 4 regions, one for each disease state. The figure also shows how a node's infectiousness varies over time, which we describe next.

The infectiousness of node v is determined by a probability density function $\beta_v : [0, \infty] \to \mathbb{R}_+$. We draw (τ_v, β_v) from a joint probability distribution P_β. This couples the duration of each disease state with its corresponding infectiousness. Further, we assume that P_β depends on the local network structure, i.e., there exists an integer $\ell > 0$ such that β_v and τ_v are drawn from $P_\beta(B_\ell(G, v))$.

Then, the epidemic progresses as follows. First, for each node v, draw β_v and τ_v from $P_\beta(B_\ell(G, v))$. Then, for each neighbor of v, draw its transmission times independently from β_v. Initially, each node is equally likely to be in $\mathcal{D}_1$, and this initialization occurs with a probability $\rho > 0$. Incorporating Algorithm 1 into this epidemic model is similar: draw β_v and τ_v for each node v from P_β, and then draw transmission times from β_v. This process yields a weighted directed graph, from which the backward edges for the next traversal can be identified.

Next, we show that our main results apply to this general epidemic model. As before, we define $\mathscr{E}_n(t)$ as the fraction of nodes in each state of the epidemic (susceptible and disease states $\mathscr{D}$) at time t in a finite graph G_n. Similarly, define $\mathscr{E}(t)$ for the epidemic state on the limit graph, and $\hat{\mathscr{E}}_{q,k,n}(t)$ as the result of the estimator with q queries and k as input.

Corollary 2.7 (Convergence of Epidemics with Time-varying Infectiousness)**.** *Let $(G_n)_{n\geq 1}$ satisfy the conditions of Theorem 2.5. Consider an epidemic model with time-varying infectiousness as above. Then the epidemic concentrates for any $t \geq 0$, $\mathscr{E}_n(t) \xrightarrow{\mathbb{P}} \mathscr{E}(t)$. Further for any given $\delta > 0$, there exists constants k, q such that, for all $t \in [0, \infty]$*

$$\limsup_{n\to\infty} \mathbb{P}\Big(|\hat{\mathscr{E}}_{q,k,n}(t) - \mathscr{E}(t)| > \delta\Big) \leq \delta. \tag{7}$$

The time-varying infection model is a dynamic and adaptable framework that allows for the incorporation of various epidemic models. As an exercise, the reader can verify how additional states, like an exposed period, can be incorporated. More generally, this model allows interventions. In this context, we highlight two applications: one that addresses vaccination and another that models social distancing.

Example 2.8 (Epidemics with Vaccination)**.** Consider a scenario where specific nodes within a network receive vaccinations based on a locally defined probability function, guaranteeing their immunity against the disease. This scenario can be represented using the time-varying infectiousness epidemic model, where $\mathscr{D} = \{I, R, V\}$ has three states of Infectious, Recovered, and Vaccinated. In this model, for vaccinated nodes, P_β allocates a β_v with zero density probability and τ_v that defines the time of vaccination at $t_3 = 0$ (as a result, $t_1 = t_2 = 0$). For all other nodes, P_β assigns an infectiousness density function as usual. ◀

Example 2.9 (Epidemics with Social Distancing)**.** The model can also apply to scenarios where individuals start practicing social distancing at a random time after becoming exposed to the disease. Hypothetically, a person might get tested and then stay home to prevent further spread. In our model, the infectiousness density, β_v, along with τ_v can be tailored to capture both the moment of infection and the duration until the adoption of social distancing measures. ◀

Remark 2.10 (Generalizations for finite graphs)**.** While the results presented in this section primarily focus on graph sequences with a local limit, it is worth noting that the findings extend to finite graphs, with similar

conditions as in tightness and stable neighborhoods that it was discussed before in Section 2.2.1. The only difference is that we need to add a notion of 'marks' (discussed in Section 3.4) to these conditions. We have chosen not to delve into these cumbersome notational adjustments to keep the primary exposition clear. ◀

2.3.2 General Starting Configurations

In our main results so far, we assumed that each node is initially infected independently with a probability $\rho > 0$. We can generalize our results to heterogeneous initial states, where a node's initial state is drawn from a probability distribution depending on its local neighborhood, for example, the node degree. As another example, the initial state could be determined by the PageRank, which can be approximated by local network structures (Garavaglia et al., 2020). To formalize this, we assume there exists a function P_ℓ, where given a node's ℓ-neighborhood $B_\ell(G, o)$ as input, $P_\ell(B_\ell(G, o))$ provides a probability distribution on the initial states of the node o, whether S, I, or R.

We further need a second condition ensuring the presence of an initially infected individual within any sufficiently long path originating from a uniformly random node; this is termed the 'locally reachable property'. To formalize this concept, let $\mathrm{Path}_r(G_n, v)$ be a uniform random path of length r among all such self-avoiding paths starting from v in graph G_n. Then $(G_n)_{n\geq 0}$ with the initial conditioned drawn from P_ℓ is locally reachable if for any $\delta > 0$,

$$\lim_{r\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\mathbb{P}\big(\mathrm{Path}_r(G_n, v) \cap I_0 = \emptyset \mid G_n\big) \geq \delta\Big) = 0, \tag{8}$$

where the inner probability is over the uniform random node v, the randomness of the path, and the initial state of infection I_0. This property obviously holds for the case of independent initial infection with probability ρ, since the chance that a path of length r does not encounter I_0 is $(1-\rho)^r$, which goes to 0 with $r \to \infty$. More generally, we have the following result:

Corollary 2.11 (General Starting Configuration)**.** *Let $(G_n)_{n\geq 1}$ satisfy the conditions of Theorem 2.5. Consider a SIR epidemic, where the starting infections are locally reachable* (8)*, and that there exists some ℓ such that the initial conditions are specified by a strictly local function P_ℓ based on ℓ-neighborhoods as defined above. Then the conclusions of Theorems 2.5 and 2.6 hold.*

Here, we consider the initial states when we define the epidemic in the limit $\mathscr{E}(t) = (s(t), i(t), r(t))$, i.e, here,

$$s(t) = \mu_\Xi(o \in \mathcal{S}(t)), \qquad i(t) = \mu_\Xi(o \in \mathcal{I}(t)), \qquad r(t) = \mu_\Xi(o \in \mathcal{R}(t)),$$

where we equipped the measure μ over rooted graphs $\mathscr{G}_\star$ with first drawing the rooted graph (G, o) and then the initial conditions P_ℓ, and denoted it as μ_Ξ. Also, similar to before, $(\mathcal{S}^{(\rho)}(t), \mathcal{I}^{(\rho)}(t), \mathcal{R}^{(\rho)}(t))$ are the sets of susceptible, infected and recovered nodes for an epidemic on (G, o) started from an initial condition that is drawn with respect to P_ℓ. See details in Section 3.4.

Example 2.12 (Application to Viral Marketing)**.** One of the practical applications of our results can be seen in the realm of viral marketing. Consider a scenario where a company wishes to maximize the spread of information about a new product. A common strategy in this context is to use a degree-based seeding approach, targeting individuals with a high degree of connectivity within a social network to initiate the spread of information. The subsequent question arises: How will information spread for a particular seeding strategy? Historically, this question has been explored for specific random graph models (see, e.g., Manshadi et al. (2020); Akbarpour et al. (2018)). By recognizing that degree-based targeting naturally satisfies the strictly local assumption required for our model, the marketing platform can apply our local estimator to predict how information will spread through the network by observing only the behavior of a few nodes. ◀

3 Preliminaries

In this section, we systematically establish a hierarchy of conditions on the local structure of graphs, each serving as an extension of its predecessor.

3.1 Graphs with Tight Neighborhood Sizes

The first condition is the notion of *tightness*, a property within a sequence of graphs that requires the number of nodes within a specific radius of a uniformly selected node to be bounded. Recall that $B_r(G, o)$ represents the subgraph of (G, o) comprising all nodes at a graph distance of at most r from o.

Definition 3.1 (Graphs with tight neighborhood sizes.)**.** Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of graphs with $|V(G_n)| = n$, and let $\varepsilon_r(G_n, k)$ be the empirical probability that $B_r(G_n, v)$ contains more than k nodes when v is chosen uniformly at random, i.e.,

$$\varepsilon_r(G_n, k) = \frac{1}{n} \sum_{v\in V(G_n)} \mathbb{1}\{|B_r(G_n, v)| > k\}. \tag{9}$$

We say that the sequence of graphs has *tight neighborhood sizes* if for all $r < \infty$ and all $\delta > 0$ there exists $k < \infty$ such that for all n large enough $\varepsilon_r(G_n, k) \leq \delta$. If the graph G_n is itself random, then we say it has tight neighborhood sizes if $\mathbb{P}(\varepsilon_r(G_n, k) \leq \delta) \geq 1 - \delta$. ◀

3.2 Graphs with Stable Neighborhood Structures

Moving forward, we expand our constraints to encompass graphs drawn from random distributions. We introduce a condition called *Stable Neighborhood Structure*. This condition ensures that the empirical distribution of local network structures is similar in different realizations of the random network.

To define this rigorously, we need to define the concept of a 'rooted graph,' which we will use in the following sections as well. A *rooted graph* is a pair (G, o) where $G = (V(G), E(G))$ is a graph with nodes in $V(G)$ and edges in $E(G)$, and $o \in V(G)$ is a specific node. The graphs (G_1, o_1) and (G_2, o_2) are *isomorphic*, denoted as $(G_1, o_1) \simeq (G_2, o_2)$, if there exists a bijection $\phi\colon V(G_1) \mapsto V(G_2)$ such that $\phi(o_1) = o_2$ and $u, v \in E(G_1)$ if and only if $\phi(u), \phi(v) \in E(G_2)$. Also, define $P_r^{(G_n)}(H^\star) = \frac{1}{|V(G_n)|} \sum_{v\in V(G_n)} \mathbb{1}\{B_r(G_n, v) \simeq H^\star\}$ as the probability that the r-ball neighborhood of a uniform random node in G_n is isomorphic to $H^\star$. For random graphs of size n, let $p_r^{(n)}(H^\star) = \mathbb{E}\big[P_r^{(G_n)}(H^\star)\big]$ represent the mean of this probability across all random realizations of G_n on graphs of size n.

Definition 3.2 (Stable Neighborhood Structure)**.** Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of (possibly random) graphs with $|V(G_n)| = n$. We say that the sequence of graphs has a *Stable Neighborhood Structure* if for all $r < \infty$, all $\delta > 0$, and any rooted graph $H^\star$, as $n \to \infty$,

$$\mathbb{P}\Big(|P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)| \geq \delta\Big) \xrightarrow{\mathbb{P}} 0. \tag{◀}$$

The stable neighborhood structure ensures that different random samples for similar-sized networks lead to the same empirical distribution of local neighborhood structures. This condition allows for the statistics of local neighborhood structures to change for different values of n. To analyze the asymptotics of graphs, a stronger requirement is needed. Not only must we ensure consistent empirical distributions for networks of similar size, but these distributions must also be asymptotically independent of the size of the graph n. This additional constraint guides us to the next concept: local convergence in probability.

3.3 Local Convergence in Probability

The foundational framework of local weak convergence was initiated independently by Aldous and Steele (2004) as well as by Benjamini and Schramm (2001). For a more comprehensive treatment, readers are directed to Bordenave (2016) or (van der Hofstad, 2024, Chapter 2).

At a high level, a sequence of graphs $(G_n)_{n\in\mathbb{N}}$ is said to exhibit local convergence if the empirical distribution governing the neighborhoods of randomly chosen nodes approximates a certain limit distribution. To define this rigorously, we must introduce a metric on the space of rooted graphs.

We denote the space of (potentially infinite) connected rooted graphs as $\mathscr{G}_\star$, where two rooted graphs are considered equivalent if they are isomorphic. Therefore, $\mathscr{G}_\star$ consists of equivalence classes of rooted graphs

modulo isomorphism. This space of rooted graphs, $\mathscr{G}_\star$, can be endowed with a metric structure denoted as d_{loc}. The metric d_{loc} between two rooted graphs (G_1, o_1) and (G_2, o_2) is defined as,

$$d_{\text{loc}}((G_1, o_1), (G_2, o_2)) = \frac{1}{1 + \inf_k\{k : B_k(G_1, o_1) \not\simeq B_k(G_2, o_2)\}}.$$

Note that this metric endows $\mathscr{G}_\star$ with the natural σ-algebra of Borel sets, allowing us in particular to consider measures μ on $\mathscr{G}_\star$.

Definition 3.3 (Local convergence in probability)**.** *Let μ be a measure on $\mathscr{G}_\star$. We define the concept of local convergence in probability for a sequence of graphs $(G_n)_{n\geq 1}$ to a limit $(G, o) \sim \mu$ as follows: For every $r \geq 0$ and $H^\star \in \mathscr{G}_\star$,*

$$\frac{1}{|V(G_n)|} \sum_{v \in V(G_n)} \mathbb{1}\{B_r(G_n, v) \simeq H^\star\} \xrightarrow{\mathbb{P}} \mu(B_r(G, o) \simeq H^\star). \tag{10}$$

This definition implies that the proportions of subgraphs in the random graph G_n converge in probability towards those prescribed by μ. The above expression is equivalent to stating that $p_r^{(G_n)}(H^\star) \xrightarrow{\mathbb{P}} \mu(B_r(G, o) \simeq H^\star)$.

Other notions of local convergence, such as local weak convergence, where the focus shifts to the convergence of expectations, and local almost sure convergence, which considers almost sure convergence, are related. However, for our current purposes, local convergence in probability is the most convenient choice, particularly due to its implication that the neighborhoods of two uniformly chosen nodes do not overlap; see (van der Hofstad, 2024, Corollary 2.18) for further details.

3.4 Mark Local Convergence

In our framework, the notion of *marks* will play a pivotal role. These marks correspond to attributes associated with the infection and recovery times of the epidemic, as well as the initial states of the nodes. We assume the marks are defined on some measurable space Ξ.

We define graph marks $\mathcal{M}(G) = ((M(v))_{v\in V(G)}, (M(v, u))_{(u,v)\in E(G)})$ to annotate G with the marks associated with both nodes and edges, where $M(v), M(v, u) \in \Xi$. Similarly, a marked rooted graph $(G, o, \mathcal{M}(G))$ is a rooted graph (G, o) with the corresponding marks. Here, edges are considered as directed with (u, v) showing the direction from u to v. This distinction is particularly relevant in our exploration of epidemics, where the traversal time along a directed edge (v, u) may differ from that of (u, v).

Given a finite graph G, we represent the probability distribution on $\mathcal{M}(G)$ as $P_\Xi(\cdot|G)$. If (G, o) is a locally finite graph with a root node o, the notation $P_\Xi(\cdot|(G, o))$ is employed. Furthermore, μ_Ξ is used to denote the measure on marked graphs in $\mathscr{G}_\star$. This is derived by initially selecting $(G, o) \sim \mu$ and subsequently assigning marks using the measure $P_\Xi(\cdot|(G, o))$. Next, we will detail how Ξ and P_Ξ are defined in the context of SIR, SIR with general starting configuration, and epidemics with time-varying infectiousness.

In the context of SIR epidemics, the mark space is denoted as $\Xi = [0, \infty] \times \{S, I, R\}$. The first component of Ξ corresponds to the transmission time for edges and the recovery time for nodes. The second component, which is relevant only to nodes, represents the initial state of the node. When defining the probability distribution $P_\Xi(\cdot|(G, o))$, the first component draws infection times for each edge from the distribution D_I, and recovery time for each node from D_R. For the second component of marks (which is independent of the first component), each node has an initial state of I with probability ρ, and if not, it is marked as S.

For epidemics with a general starting point, the space of marks remains unchanged as $\Xi = [0, \infty] \times \{S, I, R\}$. Also, the probability distribution P_Ξ on the first component corresponds to the time of transmission, and recovery stays as before. The distinction arises in the second component, representing the epidemic's initial state. In this scenario, we use a function P_ℓ that maps rooted graphs with a radius of at most ℓ to probability distributions over the states $\{S, I, R\}$. Consequently, the second component of a node's mark, $M(o)$, is drawn from $P_\ell(B_\ell(G, o))$. With this assumption, the mark of a node depends only on the ℓ-neighborhood of the root, i.e., $P_\Xi(M(o)|(G, o)) = P_\Xi(M(o)|B_\ell(G, o))$.

In the context of epidemics with time-varying infectiousness, we refine the definition of marks associated with nodes and the process of infection transmission from one node. Specifically, the node marks $(M(v))_{v\in V(G)}$ now encompass not only the initial state of node – indicating whether it is in a disease state or susceptible– but

also the density functions β_v and τ_v. Let $\mathscr{B}$ be the space of pairs of density functions such as (β_v, τ_v), where $\beta_v : [0, \infty] \to \mathbb{R}_+$ is a probability density of transmission time from v to its neighbors and $\tau_v : [t_1, \ldots, t_m] \to \mathbb{R}_+^m$ shows transition times between different disease states of the node. Then the marks on nodes take values in $\mathscr{B} \times \{0, 1\}$, with the first component identifying (β_v, τ_v) drawn from P_β, and the second component indicating whether the initial state is susceptible or disease. In addition, edge marks, represented as $M(v, u)$, are drawn independently from β_v, indicating the transmission time from v to u. So, the space of marks for edges is $\mathbb{R}_+$. As a result, the space of marks on the graph is the union of node marks and edge marks $\Xi = \mathscr{B} \times \{0, 1\} \cup \mathbb{R}_+$. Further, P_Ξ is defined by first drawing node marks and then edge marks as described above.

4 Proof Outline

In this section, we outline the main ideas of the proofs. We employ a second-moment argument for finite graphs, demonstrating that truncating the epidemic at a constant radius approximates the epidemic on the entire graph. Details for deterministic graphs can be found in Section 4.1 and for random graphs in Section 4.2. We then incorporate local convergence in Section 4.3. Lastly, Section 4.4 addresses generalizations via marked graph convergence.

4.1 Proofs for Finite Deterministic Graphs

We start by proving that running the epidemic within the r-ball of each node, rather than across the entire graph, results in an outcome that concentrates around the true time evolution of epidemics. This idea is equivalent to running Algorithm 1 for n queries with each node as a starting point, then selecting a sufficiently large k to cover the r-ball of each node. Using a second-moment argument, we will bound both the mean and variance of the truncation.

To formalize our approach, we introduce the notation $T^{(r)}$. This function maps a rooted marked graph $(G, o, \mathcal{M}(G))$ to the infection time of o under the assumption that the network is confined to $B_r(G, o)$. Formally, $T^{(r)}$ assigns a non-negative number to a rooted marked graph $(G, o, \mathcal{M}(G))$, which is equal to the length of the shortest path from the set of initially infected nodes in $B_r(G, o)$ to the root o (with respect to the weighted distance $\mathrm{dist}_{(T)}(\cdot, \cdot)$ defined in Section 2). When the graph is given in the context, we use $T^{(r)}(o)$ as a short form of $T^{(r)}(G, o, \mathcal{M}(G))$.

Our goal is to approximate $S_n^{(\rho)}(t)$ by a sum of functions defined on balls of radius r, namely

$$S_{n,r}^{(\rho)}(t) = \frac{1}{n} \sum_{v \in V(G_n)} \mathbb{1}\{t < T^{(r)}(v)\}. \tag{11}$$

Similarly one can define $I_{n,r}^{(\rho)}(t)$, $R_{n,r}^{(\rho)}(t)$, and the vector $\mathscr{E}_{n,r}^{(\rho)}(t) = (S_{n,r}^{(\rho)}(t), I_{n,r}^{(\rho)}(t), R_{n,r}^{(\rho)}(t))$. The following two lemmas bound the first and second moment of this truncation:

Lemma 4.1 (Local Approximation - First Moment)**.** *For a given finite graph G_n,*

$$\mathbb{E}_\Xi \left[\sup_{t \geq 0} \left| \mathscr{E}_n^{(\rho)}(t) - \mathscr{E}_{n,r}^{(\rho)}(t) \right| \right] \leq (1 - \rho)^r,$$

where the expectation is with respect to the epidemic and the random set of initially infected nodes (or, equivalently, the marks on G_n).

The proof is based on showing that the shortest path, in terms of $\mathrm{dist}_{(T)}$, from a node to the set of initially infected nodes traverses only a limited number of graph nodes. This is because the initially infected nodes are 'locally reachable', and the likelihood of not encountering them reduces geometrically (see Appendix A.1).

Subsequently, we bound the variance of this approximation for the second moment. The main idea is that the local epidemic processes (which define $T^{(r)}$) for nodes separated by more than distance $2r$ are independent. Consequently, we can constrain the variance by counting node pairs separated by over-distance r. This quantity is denoted as

$$\varepsilon_r(G_n) = \frac{1}{n^2} \left| \{(x, y) \in V(G_n) \times V(G_n) \colon \mathrm{dist}_{G_n}(x, y) \leq r\} \right|,$$

where $\text{dist}_{G_n}(x, y)$ is the graph distance of x and y in G_n. The detail of the proof appears in Appendix A.2. Later, we provide bounds on $\varepsilon_r(G_n)$ based on $\varepsilon_r(G_n, k)$, which was defined in the statement of Theorem 2.1.

Lemma 4.2 (Local Approximation - Second Moment)**.** *Let G_n be a deterministic graph with n nodes. Then*

$$\sup_{t\geq 0} \text{Var}_{\Xi}(S^{(\rho)}_{n,r}(t)) \leq \frac{1}{n} + \varepsilon_{2r}(G_n),$$

where the variance is over the randomness of the epidemic process.

Note that the function $S^{(\rho)}_{n,r}$ is effectively equivalent to the estimator $\hat{S}^{(\rho)}_{n,r,n}$ when making n queries with an input of r to Algorithm 1. Further, Lemmas 4.1 and 4.2 provide the foundation for establishing the concentrations of $S^{(\rho)}_{n,r}(t)$ and $S^{(\rho)}_{n}(t)$. To conclude the proof of Theorem 2.1, the main step is to determine the accuracy of q queries $\hat{S}^{(\rho)}_{q,r,n}$ to approximate $S^{(\rho)}_{n,r}$. For this purpose, first, we condition on the epidemic process to bound the error of choosing q using standard concentration arguments. Then, to get the overall accuracy of the estimator, we use the variance bound in Lemma 4.2 to control the estimator's value across different realizations of the epidemic process. See Appendix A.3.

Theorem 2.1 provides explicit bounds on the estimator's error for a given graph. This bound can be made as narrow as desired for graphs that meet the 'tightness' condition in Definition 3.1. The following theorem formalizes this – even with a constant number of queries – the estimator closely match the true time evolution of epidemics, $(S^{(\rho)}_n(t), I^{(\rho)}_n(t), R^{(\rho)}_n(t))$, to any chosen precision:

Theorem 4.3 (Local Estimation of Tight Graphs)**.** *Let $(G_n)_{n\in\mathbb{N}}$ be a sequence of graphs with tight neighborhood sizes. Then $\left|\mathscr{E}^{(\rho)}_n(t) - \mathbb{E}[\mathscr{E}^{(\rho)}_n(t)]\right| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. Furthermore, given any $\delta > 0$, there exists constants N, q, k such that for any $n > N$ and $t \geq 0$,*

$$\mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{q,k,n}(t) - \mathscr{E}^{(\rho)}_n(t)| > \delta\Big) \leq \delta.$$

In Appendix A.5, we present an example emphasizing the importance of the tightness condition, illustrating that in structures like the star graph, even the final infection size does not concentrate.

4.2 Proofs for Finite Random Graphs

The proof largely mirrors that of deterministic graphs, employing a similar second-moment argument. For the first moment, Lemma 4.1 suffices as we can use linearity of expectation to get convergence of first-moment with the randomness of G_n taken into account. However, for bounding the variance as in Lemma 4.2, we must extend our result to accommodate the randomness of the neighborhood structure. Definition 3.2 ensures minimal variance between the expected time evolution of epidemics across different network realizations as it is highlighted in Lemma 4.4. See Appendix A.2. Then, the proof of the Theorem 2.3 follows the exact steps of Theorem 2.1, which appears in Appendix B.

Lemma 4.4 (Local Approximation for Random Networks - Second Moment)**.** *Let $(G_n)_{n\geq 1}$ be a sequence of (possibly random) graphs with tight and stable neighborhood structures (see Definitions 3.1 and 3.2). Then for any given $\delta > 0$, and large enough n,*

$$\sup_{t\geq 0} \text{Var}(S^{(\rho)}_{n,r}(t)) \leq \delta,$$

where the variance is over the randomness of the epidemic process and the graph G_n.

4.3 Proofs for Growing Graphs

We establish the local convergence of the epidemic in three primary stages. First, as in Theorem 2.3, we prove that an epidemic restricted to a constant radius ball around nodes (represented as $\mathscr{E}_{n,r}(t)$ in (11)) concentrates around the epidemic spanning the entire graph $\mathscr{E}_n(t)$. The second step is the local approximation of the limit graph with a similar truncation. This approach mirrors the method used for the finite graph, which we detail in Lemma 4.5. The final stage (Lemma 4.6) ensures the convergence of the truncated epidemic in the finite graph to that of the limit graph.

Starting with the local approximation of the epidemic in the limit, recall the definition of $T^{(k)}(o, G, \mathcal{M}(G))$ from Section 4.1. For the limit graph, define $s_k(t) = \mu_\Xi\Big(\mathbb{1}\{t < T^{(k)}(G, o, \mathcal{M}(G))\}\Big)$. Similarly, $i_k(t)$, and $r_k(t)$ can be defined. Then we can extend Lemma 4.1 for the limit graph. The proof follows similar steps as in the proof of Lemma 4.1 followed by monotone convergence.

Lemma 4.5 (Local Approximation of the Limit)**.** *For any (deterministic) measure μ on $\mathcal{G}\star$, and any integers k and k',*

$$\mu_\Xi\left[\sup_{t\geq 0}\Big|\mathbb{1}\{t < T^{(k)}(G, o, \mathcal{M}(G))\} - \mathbb{1}\{t < T^{(k')}(G, o, \mathcal{M}(G))\}\Big|\right] \leq (1-\rho)^{\min\{k,k'\}}.$$

Thus, $s(t) = \lim_{k\to\infty} s_k(t)$, $i(t) = \lim_{k\to\infty} i_k(t)$ and $r(t) = \lim_{k\to\infty} r_k(t)$ are well-defined, and

$$\sup_{t\geq 0}|(s_k(t), i_k(t), r_k(t)) - (s(t), i(t), r(t))| \leq (1-\rho)^k.$$

Next, we will show that $S^{(\rho)}_{n,r}(t)$ is local, meaning that it converges to $s_r(t)$ uniformly in t. This step is essential since $S^{(\rho)}_{n,r}$ is not a continuous function of t with n discontinuities at the time of infection of each node. The proof is based on conditioning on the structure of the graph in the local neighborhood and then using the tightness of graphs with local limits to bound the probability.

Lemma 4.6 (Convergence of Local Approximation)**.** *Let $(G_n)_{n\geq 1}$ be a graph sequence that converges locally in probability to $(G, o) \sim \mu$, and let $(G_n, \mathcal{M}(G_n))$ be the marked graph. Then for any $t \in [0, \infty]$, $\mathbb{E}_\Xi[S^{(\rho)}_{n,r}(t)] \xrightarrow{\mathbb{P}} s_r(t)$, where the convergence in probability is with respect to the randomness of G_n.*

Now, to prove Theorem 2.5, first we can apply Lemmas 4.1 and 4.4 to prove that $S^{(\rho)}_{n,r}(t)$ is a good approximation of $S^{(\rho)}_n(t)$. Then we can subsequently apply Lemma 4.6 and then Lemma 4.5 to prove convergence of $S^{(\rho)}_n(t)$ to $s(t)$ uniformly in t. Finally, Theorem 2.6 is an immediate corollary of Theorems 2.3 and 2.5, as convergent graphs satisfy both tightness and stable neighborhood conditions. The details appear in Appendix C.

4.4 Generalizations

4.4.1 Proofs for General Epidemics

We start with Corollary 2.7, which follows very similar steps as in the proof of Theorems 2.5 and 2.6, with a small subtlety. Before, it was enough to prove the concentrations for the number of susceptible nodes, and that naturally led to concentration for recovered and, subsequently, infectious nodes. Now, we need to stretch this idea a bit. We will show that the number of nodes in states S, or $\mathcal{D}_1$ through $\mathcal{D}_i$ are concentrated. Then, the proof works similarly to before; effectively, you can think of nodes in the union of these $i+1$ states as a new susceptible state. Once we nail down the convergence for these combined states, the convergence of nodes in a specific state like $\mathcal{D}_i$ comes into focus by subtracting from these larger unions of two of these unions. See the details in Appendix D.1.

4.4.2 Proofs for General Starting Configuration

Again, the main idea is to truncate the epidemic to a constant radius r and argue that this offers a good approximation of the epidemic on the entire graph. As before, we can truncate the disease at a finite distance r, and show it concentrates both at the limit and for finite graphs. The proof of concentration, as before, is by a second-moment argument. We can generalize the first-moment Lemma 4.5 to the case that the initial condition is locally reachable. Further, the bound on the second moment is a direct implication of Lemma 4.4, since the truncated epidemic of two nodes that are at a large distance are still independent. The only caveat is that we need to add the distance ℓ, which determines the radius that the starting configuration P_ℓ depends on. See Appendix D.2.

5 Experiments

To empirically validate the asymptotic result derived in Theorems 2.3 and 4.3, we ran experiments using both synthetic and real-world networks. The aim of our investigation is to assess the applicability and accuracy of the proposed estimator (1) in predicting the time evolution and final infection size of epidemic outbreaks in practice.

Synthetic Networks: We generated synthetic networks using two well-established models: Preferential Attachment (Barabási and Albert, 1999) and Random Geometric Graphs (Gilbert, 1959). The Preferential Attachment model represents the growth of scale-free networks, where new nodes join the network and preferentially connect to existing nodes based on their degrees. In our experiments, we set the parameter $m = 3$, indicating that each new node forms exactly three connections with existing nodes.

For Random Geometric Graphs, nodes are randomly distributed in a Euclidean space with their positions defined by x and y axes drawn uniformly at random from $[0, \sqrt{n}]$, where n denotes the size of the graph. The connection radius for the Random Geometric Graph is set to 1.5, ensuring an average degree of approximately 7.06 as $n \to \infty$. For both models, we created synthetic networks of varying sizes, ranging from 500 to $10,000$ nodes.

Copenhagen Interaction Network: To further validate the estimator's effectiveness in a real-world setting, we utilized a temporal network from the Copenhagen Networks Study (Sapiezynski et al., 2019). This dataset recorded interactions among university students at 5-minute intervals over a four-week duration. The interactions were based on Received Signal Strength Indicator (RSSI) values from Bluetooth signal strength measurements, reflecting the physical proximity between individuals.

We focused on a specific 12-hour interval (from 8 am to 8 pm) on the fourth day of the study. Within this timeframe, we considered only those connections that were within a distance of 6 feet or an equivalent RSSI value of -74.25, ensuring that we concentrate on significant interactions representative of close proximity encounters. The resulting graph has 422 nodes with an average degree of 7.89, reflecting a moderate level of connectivity among the participants. The degree distribution of the network is illustrated in Figure 6.

SafeGraph San Francisco Network: To extend our investigation to a different real-world scenario, we utilized the mobility dataset for San Francisco County. This data set is derived by (Chang et al., 2021) from SafeGraph data and is structured as a bipartite network with time-varying edges. This bipartite network has dynamic edges between Census Block Groups (CBG) — geographic units of 600 to 3,000 people — and Points of Interest (POI). Edge weights represent the number of CBG visitors to a POI in a given hour.

To construct our edge-weighted network, we aggregated the mobility data for San Francisco County over six hours on March 1, 2020, specifically from 6:00 am to 12:00 pm. This aggregation process resulted in a comprehensive representation of the interactions between $28,713$ POIs and $2,943$ CBGs, yielding a total of $31,656$ nodes and $82,022$ weighted edges. See Figure 7 for weighted degree distribution. For our analysis, we treated each POI and CBG as individual nodes without accounting for the population of CBGs. The edge weights were used as the transmission rates between the nodes.

Experiment Details and Epidemic Parameters: In our experiments, the infection and recovery times were drawn from exponential random variables with rates set to 1. The initial infection probability for a node was set to 0.01. For the estimator (1), we used a total of 10 queries ($q = 10$) and varied the input budget (k) in Algorithm 1, selecting values in the range of 2 to 9. A minor difference in the implementation of Algorithm 1 from its description in the paper was introduced: once the backward process identifies k nodes and if none are initially infected, we randomly select one node from these k nodes that are furthest (in terms of graph distance) from the root to be initially infected [1] We repeated the experiment for $1,000$ times to evaluate confidence intervals on the estimator's accuracy.

[1] Introducing this modification led to faster convergence in practice. Because with the original algorithm, the likelihood of encountering an initially infected node among a sample of, say, 5 nodes was not substantial, with probabilities calculated as $1 - .9^5 \approx .40$. Although the proofs remain valid with this modification, we opted to retain the original algorithm description in the paper for simplicity and to avoid discussing numerous variations.

Performance Evaluation: To evaluate the estimator's performance, we conducted 1,000 simulations for each network and each choice of k. For comparison, we ran the SIR process with the same infection rate, recovery rate, and initial infection probability over 1,000 iterations, using the Epidemic on Network package in Python (Miller and Ting, 2020; Kiss et al., 2017). The average of these 1,000 runs was considered the ground-truth time evolution. See Figures 2 to 5. Throughout our assessment, distinct scenarios unfolded. For the Copenhagen dataset and the Preferential Attachment Model, an input budget of $k = 6$ for each query sufficed. In the context of San Francisco and Geometric Random Graphs, a budget of $k = 9$ proved optimal. Given the fixed number of queries set at 10, this equates to utilizing merely 60 nodes in the former cases and 90 nodes in the latter instances. Remarkably, this translated to using only 0.28% of San Francisco, 14.9% of Copenhagen, 0.9% of Random Geometric Graphs, and 0.6% of Preferential Attachment nodes, yet maintaining highly accurate predictions of epidemics.

Furthermore, to assess the similarity between the estimated time evolution and the ground-truth, we computed the Euclidean distance and Pearson correlation of the estimated time series with the ground-truth time series. The results are presented in Tables 1 to 4. We also considered how the size of the graph would affect the estimator's accuracy in Table 5.

6 Conclusion

In this research, we developed a novel approach to understanding the intricate dynamics of epidemics on diverse network structures. Through the introduction of a local estimator and its robust theoretical guarantees, our result shows the inherent 'local' nature of epidemic behaviors even in large networks. Notably, our empirical results validate the precision of our estimator on various datasets.

Our results carry profound policy implications. As the world faces recurring epidemics, our findings highlight the advantages of targeted data-gathering strategies. Rather than obtaining vast amounts of data indiscriminately, policymakers and researchers may consider leaning on local network structures, ensuring both efficiency in collection and enhanced predictive accuracy. Moreover, our algorithm can be used for fast implementation of intervention strategies on networks, enabling the comparison of different mitigations. In the past, many relied on mean-field models or specific random network models for this purpose (Birge et al., 2022; Acemoglu et al., 2023).

In our work, we bring the theory of local convergence into the realm of operations research. Given the vast expanse of problems governed by network dynamics, numerous applications stand to benefit from our approach. The diffusion of misinformation (Mostagir and Siderius, 2023), the complexities of viral marketing (Ho et al., 2002; Manshadi et al., 2020), pricing for accelerating diffusion (Kalish and Lilien, 1983; Shen et al., 2011), network learning (Hu et al., 2019), and many other studies have traditionally been restricted by assumptions about network models. However, our method offers a fresh lens, providing a deeper understanding without necessitating network model assumptions.

Acknowledgements. This research was partially conducted during the visit of all authors to the Simons Institute of Theoretical Computing in the "Graph Limits and Processes on Networks: From Epidemics to Misinformation" program in the fall of 2022. The work of RvdH is supported in parts by the NWO through the Gravitation NETWORKS grant 024.002.003.

References

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2023). Testing, voluntary social distancing, and the spread of an infection. *Operations Research*.

Ajorlou, A., Jadbabaie, A., and Kakhbod, A. (2018). Dynamic pricing in social networks: The word-of-mouth effect. *Management Science*, 64(2):971–979.

Akbarpour, M., Malladi, S., and Saberi, A. (2018). Diffusion, seeding, and the value of network information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 641–641.

Aldous, D. and Steele, J. (2004). The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin.

Alimohammadi, Y., Borgs, C., and Saberi, A. (2022). Algorithms using local graph features to predict epidemics. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3430–3451. SIAM.

Alimohammadi, Y., Borgs, C., and Saberi, A. (2023). Locality of random digraphs on expanders. *The Annals of Probability*, 51(4):1249–1297.

Amini, H. and Minca, A. (2016). Inhomogeneous financial networks and contagious links. *Operations Research*, 64(5):1109–1120.

Andreasen, V., Lin, J., and Levin, S. A. (1997). The dynamics of cocirculating influenza strains conferring partial cross-immunity. *Journal of Mathematical Biology*, 35:825–842.

Aparicio, J. P., Capurro, A. F., and Castillo-Chavez, C. (2000). Transmission and dynamics of tuberculosis on generalized households. *Journal of Theoretical Biology*, 206(3):327–341.

Baek, J., Farias, V. F., Georgescu, A., Levi, R., Peng, T., Sinha, D., Wilde, J., and Zheng, A. (2021). The limits to learning a diffusion model. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 130–131.

Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one*, 6(1):e16591.

Ball, F. and Britton, T. (2022). Epidemics on networks with preventive rewiring. *Random Structures & Algorithms*, 61(2):250–297.

Ball, F., Sirl, D., and Trapman, P. (2010). Analysis of a stochastic sir epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73.

Bampo, M., Ewing, M. T., Mather, D. R., Stewart, D., and Wallace, M. (2008). The effects of the social structure of digital networks on viral marketing performance. *Information Systems Research*, 19(3):273–290.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144):1236498.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Bartlett, M. (1949). Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):211–229.

Bass, F. M. (1969). A new product growth for model consumer durables. *Management science*, 15(5):215–227.

Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjichristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D., and Tsiodras, S. (2021). Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, 599(7883):108–113.

Benjamini, I. and Schramm, O. (2001). Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, **6**:no. 23, 13 pp. (electronic).

Berger, N., Borgs, C., Chayes, J., and Saberi, A. (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *Annals of Probability*, **42**(1):1–40.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalite causee par la petite verole, et des avantages de l'inoculation pour la prevenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45.

Bertsimas, D., Boussioux, L., Cory-Wright, R., Delarue, A., Digalakis, V., Jacquillat, A., Kitane, D. L., Lukin, G., Li, M., Mingardi, L., et al. (2021). From predictions to prescriptions: A data-driven response to COVID-19. *Health Care Management Science*, 24:253–272.

Bhamidi, S., van der Hofstad, R., and Komjáthy, J. (2014). The front of the epidemic spread and first passage percolation. *Journal Of Applied Probability*, 51(A):101–121.

Birge, J. R., Candogan, O., and Feng, Y. (2022). Controlling epidemic spread: Reducing economic losses with targeted closures. *Management Science*, 68(5):3175–3195.

Blower, S. M., Mclean, A. R., Porco, T. C., Small, P. M., Hopewell, P. C., Sanchez, M. A., and Moss, A. R. (1995). The intrinsic transmission dynamics of tuberculosis epidemics. *Nature Medicine*, 1(8):815–821.

Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316.

Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, **31**(1):3–122.

Bordenave, C. (2016). Lecture notes on random graphs and probabilistic combinatorial optimization. Version April 8, 2016. Available at `http://www.math.univ-toulouse.fr/~bordenave/coursRG.pdf`.

Britton, T., Pardoux, E., Ball, F., Laredo, C., Sirl, D., and Tran, V. C. (2019). *Stochastic epidemic models with inference*, volume 2255. Springer.

Budhiraja, A., Dupuis, P., and Fischer, M. (2012). Large deviation properties of weakly interacting processes via weak convergence methods. *Annals of Probability*, 40(1):74–102.

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., and Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87.

Chen, J., Hoops, S., Marathe, A., Mortveit, H., Lewis, B., Venkatramanan, S., Haddadan, A., Bhattacharya, P., Adiga, A., Vullikanti, A., et al. (2022). Effective social network-based allocation of COVID-19 vaccines. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4675–4683.

Chin, A., Eckles, D., and Ugander, J. (2022). Evaluating stochastic seeding strategies in networks. *Management Science*, 68(3):1714–1736.

Coppini, F., Dietert, H., and Giacomin, G. (2020). A law of large numbers and large deviations for interacting diffusions on erdős–rényi graphs. *Stochastics and Dynamics*, 20(02):2050010.

Cvjetanović, B., Grab, B., and Uemura, K. (1971). Epidemiological model of typhoid fever and its use in the planning and evaluation of antityphoid immunization and sanitation programmes. *Bulletin of the World Health Organization*, 45(1):53.

Decreusefond, L., Dhersin, J.-S., Moyal, P., and Tran, V. C. (2012). Large graph limit for an sir process in random network with heterogeneous connectivity. *The Annals of Applied Probability*, 22(2):541 – 575.

Dembo, A. and Montanari, A. (2010). Gibbs measures and phase transitions on sparse random graphs. *Braz. J. Probab. Stat.*, **24**(2):137–211.

Dimitrov, N. B. and Meyers, L. A. (2010). Mathematical approaches to infectious disease prediction and control. In *Risk and Optimization in an Uncertain World*, pages 1–25. INFORMS.

Drakopoulos, K. and Zheng, F. (2017). Network effects in contagion processes: Identification and control. *Columbia Business School Research Paper*, (18-8).

Eckles, D., Esfandiari, H., Mossel, E., and Rahimian, M. A. (2022). Seeding with costly network information. *Operations Research*, (4):2318–2348.

Erdős, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60.

Eubank, S., Guclu, H., Anil Kumar, V., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184.

Feder, G. and Umali, D. L. (1993). The adoption of agricultural innovations: a review. *Technological forecasting and social change*, 43(3-4):215–239.

Ford, E. W., Menachemi, N., and Phillips, M. T. (2006). Predicting the adoption of electronic health records by physicians: when will health care be paperless? *Journal of the American Medical Informatics Association*, 13(1):106–112.

Ganguly, A. and Ramanan, K. (2022). Hydrodynamic limits of non-markovian interacting particle systems on sparse graphs. *arXiv preprint arXiv:2205.01587*.

Garavaglia, A., Hazra, R., van der Hofstad, R., and Ray, R. (2022). Universality of the local limit in preferential attachment models. arXiv:2212.05551 [math.PR].

Garavaglia, A., van der Hofstad, R., and Litvak, N. (2020). Local weak convergence for pagerank.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Goel, S., Anderson, A., Hofman, J., and Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, 62(1):180–196.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660.

Gupta, S., Hill, A., Kwiatkowski, D., Greenwood, A. M., Greenwood, B. M., and Day, K. P. (1994). Parasite virulence and disease patterns in plasmodium falciparum malaria. *Proceedings of the National Academy of Sciences*, 91(9):3715–3719.

Gupta, S., Starr, M. K., Farahani, R. Z., and Asgari, N. (2022). Om forum—pandemics/epidemics: Challenges and opportunities for operations management research. *Manufacturing & Service Operations Management*, 24(1):1–23.

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K. T., Edmunds, W. J., Frost, S. D., Funk, S., et al. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, 347(6227):aaa4339.

Ho, T.-H., Savin, S., and Terwiesch, C. (2002). Managing demand and sales dynamics in new product diffusion under supply constraint. *Management science*, 48(2):187–206.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Hu, M. M., Yang, S., and Xu, D. Y. (2019). Understanding the social learning effect in contagious switching behavior. *Management Science*, 65(10):4771–4794.

Isham, V. (1988). Mathematical modelling of the transmission dynamics of hiv infection and aids: a review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 151(1):5–30.

Jackson, M. O. and Yariv, L. (2005). Diffusion on social networks. *Economie Publique (Public Economics)*, 16(1):3–16.

Jacobsen, K. A., Burch, M. G., Tien, J. H., and Rempała, G. A. (2016). The large graph limit of a stochastic epidemic model on a dynamic multilayer network. *arXiv preprint arXiv:1605.02809*.

Janson, S., Luczak, M., and Windridge, P. (2014). Law of large numbers for the sir epidemic on a random graph with given degrees. *Random Structures & Algorithms*, 45(4):726–763.

Kalish, S. and Lilien, G. L. (1983). Optimal price subsidy policy for accelerating the diffusion of innovation. *Marketing Science*, 2(4):407–420.

Kaplan, E. H. (2020). Om forum—COVID-19 scratch models to support local decisions. *Manufacturing & Service Operations Management*, 22(4):645–655.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

Kim, L., Abramson, M., Drakopoulos, K., Kolitz, S., and Ozdaglar, A. (2014). Estimating social network structure and propagation dynamics for an infectious disease. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings 7*, pages 85–93. Springer.

Kiss, I. Z., Miller, J. C., Simon, P. L., et al. (2017). Mathematics of epidemics on networks. *Cham: Springer*, 598.

Komjáthy, J. and Lodewijks, B. (2020). Explosion in weighted hyperbolic random graphs and geometric inhomogeneous random graphs. *Stochastic Process. Appl.*, **130**(3):1309–1367.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Phys. Rev. E (3)*, 82(3):036106, 18.

Kurauskas, V. (2022). On local weak limit and subgraph counts for sparse random graphs. *Journal of Applied Probability*, 59(3):755–776.

Lacker, D., Ramanan, K., and Wu, R. (2019). Local weak convergence for sparse networks of interacting processes. *arXiv preprint arXiv:1904.02585*.

Larson, R. C. (2007). Simple models of influenza progression within a heterogeneous population. *Operations research*, 55(3):399–412.

Lashari, A. A., Serafimović, A., and Trapman, P. (2021). The duration of a supercritical sir epidemic on a configuration model. *Electronic Journal of Probability*, 26:1–49.

Lee, Y.-J., Hosanagar, K., and Tan, Y. (2015). Do i follow my friends or the crowd? information cascades in online movie ratings. *Management Science*, 61(9):2241–2258.

Lloyd-Smith, J. O., George, D., Pepin, K. M., Pitzer, V. E., Pulliam, J. R., Dobson, A. P., Hudson, P. J., and Grenfell, B. T. (2009). Epidemic dynamics at the human-animal interface. *science*, 326(5958):1362–1367.

Lobel, I., Sadler, E., and Varshney, L. R. (2017). Customer referral incentives and social media. *Management Science*, 63(10):3514–3529.

Mamani, H., Chick, S. E., and Simchi-Levi, D. (2013). A game-theoretic model of international influenza vaccination coordination. *Management Science*, 59(7):1650–1670.

Mandal, S., Sarkar, R. R., and Sinha, S. (2011). Mathematical models of malaria-a review. *Malaria Journal*, 10(1):1–19.

Manshadi, V., Misra, S., and Rodilitz, S. (2020). Diffusion in random networks: Impact of degree distribution. *Operations research*, (6):1722–1741.

May, R. M. and Anderson, R. M. (1987). Commentary transmission dynamics of hiv infection. *Nature*, 326(137):10–1038.

Mihara, S., Tsugawa, S., and Ohsaki, H. (2015). Influence maximization problem for unknown social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1539–1546.

Miller, J. C. and Ting, T. (2020). Eon (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *arXiv preprint arXiv:2001.02436*.

Mostagir, M. and Siderius, J. (2023). Social inequality and the spread of misinformation. *Management Science*, 69(2):968–995.

Mukherjee, U. K. and Seshadri, S. (2022). Epidemic modeling, prediction, and control. In *Tutorials in Operations Research: Emerging and Impactful Topics in Operations*, pages 1–35. INFORMS.

Netrapalli, P. and Sanghavi, S. (2012). Learning the graph of epidemic cascades. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, page 211–222, New York, NY, USA. Association for Computing Machinery.

Ross, R. and Hudson, H. P. (1917). An application of the theory of probabilities to the study of a priori pathometry.—part iii. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 93(650):225–240.

Sapiezynski, P., Stopczynski, A., Lassen, D. D., and Lehmann, S. (2019). Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):315.

Scarpino, S. V. and Petri, G. (2019). On the predictability of infectious disease outbreaks. *Nature Communications*, 10(1):898.

Shen, W., Duenyas, I., and Kapuscinski, R. (2011). New product diffusion decisions under supply constraints. *Management Science*, 57(10):1802–1810.

Stein, S., Eshghi, S., Maghsudi, S., Tassiulas, L., Bellamy, R. K., and Jennings, N. R. (2017). Heuristic algorithms for influence maximization in partially observable social networks. In *SocInf@ IJCAI*, pages 20–32.

Trapman, P. (2007). On analytical approaches to epidemics on networks. *Theoretical population biology*, 71(2):160–173.

van der Hofstad, R. (2024). Random graphs and complex networks. Vol. 2. In preparation, see `http://www.win.tue.nl/~rhofstad/NotesRGCNII.pdf`.

van der Hofstad, R., Hoorn, P. v. d., and Maitra, N. (2023). Local limits of spatial inhomogeneous random graphs. *Advances in Applied Probability*, pages 1–48.

van der Hofstad, R., Komjáthy, J., and Vadon, V. (2021). Random intersection graphs with communities. *Adv. in Appl. Probab.*, 53(4):1061–1089.

van der Hofstad, R., van Leeuwaarden, J. S., and Stegehuis, C. (2015). Hierarchical configuration model. *arXiv preprint arXiv:1512.08397*.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270.

Wilder, B., Immorlica, N., Rice, E., and Tambe, M. (2018). Maximizing influence in an unknown social network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wu, J. T., Wein, L. M., and Perelson, A. S. (2005). Optimization of influenza vaccine selection. *Operations Research*, 53(3):456–476.

Yang, Y., Nishikawa, T., and Motter, A. E. (2017). Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365):eaan3184.

Yule, G. U. (1925). Ii.—a mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213(402-410):21–87.

A Concentration of Epidemic - Proof Details

A.1 Local Approximation - Proof of Bounds on the First Moment

In this section, we delve into two closely related proofs, both of which focus on the concentration of the first moment of a local approximation. These proofs leverage the inherent structure of our models and the local reachability of nodes to shed light on the nuances of approximation within the confines of the given parameters.

The first proof, corresponding to Lemma 4.1, shows how the path to an initial infection can be bounded within a specific radius due to the fact that initially infected nodes are 'locally reachable'. The second proof (Lemma 4.5) builds upon the foundation set by the first. By concentrating on the characteristics of the nodes and their infection times, we derive insights into the monotonic behaviors of our models and how they converge in specific scenarios.

Proof of Lemma 4.1. The proof is a standard argument showing that the shortest path (in terms of $\mathrm{dist}_{(T)}$) from a node to its initial infection can be constrained within a bounded radius due to the fact that the initially infected nodes are 'locally reachable.'

We first draw the marks corresponding to the contact and recovery times, and we only keep the initial infection random. Recall the notations of $T^{(r)}$ which maps a rooted marked graphs $(G, o, \mathcal{M}(G))$ to the infection time of o under the assumption that the network is restricted to $B_r(G, o)$. Using this, $T^{(\infty)}(v)$ refers to the actual infection time of v in G_n. Note that the difference in $S_n(t)$ and $S_{n,r}(t)$ is in the set of nodes for which $T^{(\infty)}(v) \neq T^{(r)}(v)$, i.e.,

$$\mathbb{E}_{\Xi}\left[\sup_{t\geq 0}\left|\mathscr{E}_n^{(\rho)}(t) - \mathscr{E}_{n,r}^{(\rho)}(t)\right|\right] \leq \mathbb{E}_{\Xi}\left[\frac{1}{n}\left|\{v : T^{(\infty)}(v) \neq T^{(r)}(v)\}\right|\right] = \mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\Big),$$

where o_n is a uniformly random chosen node from $V(G_n)$. Consider the shortest path that identifies the infection time $T^{(\infty)}(v)$. If $T^{(\infty)}(v) \neq T^{(r)}(v)$ then the graph length of this shortest path is at least $r+1$, otherwise the infection time was already identified in the r-neighborhood. Moreover, the initial r nodes of this shortest path toward determining $T^{(\infty)}(v)$ should not contain any initially infected node. If it did, the path would reach another initially infected node faster. Therefore,

$$\mathbb{P}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n) \mid \mathcal{M}(G_n)\Big) \leq (1-\rho)^r,$$

where the probability is with respect to the initial infections. Now, if we take the probability over the marks of epidemic, we get that

$$\mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\Big) \leq (1-\rho)^r,$$

which in turn provides an upper bound for $\mathbb{E}_{\Xi}\left[\frac{1}{n}\left|\{v : T^{(\infty)}(v) \neq T^{(r)}(v)\}\right|\right]$. □

Proof of Lemma 4.5. The first part of the lemma follows from the same argument as in the proof of Lemma 4.1. Then by noting that $s_k(t)$ is monotone decreasing in k, the limit $s(t) = \lim_{k\to\infty} s_k(t)$ is well-defined, and the bound on their difference follows from the first part. Similarly, $r_k(t)$ is monotone increasing in k, and the same argument holds. Finally, we can finish the argument by noting that $i_k(t) = 1 - s_k(t) - r_k(t)$. □

A.2 Local Approximation - Proof of Bounds on the Second Moment

In this section, we delve into the second moment of our local approximation. The essence lies in understanding the interplay between node distances and their implications for the variance of our local approximation.

Our first proof, Lemma 4.2, highlights the independence of events for nodes that are sufficiently far apart in the graph, allowing us to derive a precise bound for the variance of $S_{n,r}^{(\rho)}(t)$.

Proof of Lemma 4.2. The proof follows from the fact that the events $\{t < T^{(r)}(x)\}$ and $\{t < T^{(r)}(y)\}$ are independent if $\text{dist}_{G_n}(x, y) > 2r$. Let $Z_v(t) = \mathbb{1}\{t < T^{(r)}(G_n, v, \mathcal{M}(G_n))\}$. Then, we can write

$$\begin{aligned}
n^2\text{Var}(S_{n,r}^{(\rho)}(t)) &= \mathbb{E}\Big[\big(\sum_{v\in V(G_n)} Z_v(t) - \sum_{v\in V(G_n)} \mathbb{E}(Z_v(t))\big)^2\Big] \\
&= \sum_{v\in V(G_n)} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)^2\Big] + \sum_{\text{dist}_{G_n}(v,u)\le 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] \\
&\quad + \sum_{\text{dist}_{G_n}(v,u)> 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] \\
&= \sum_{v\in V(G_n)} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)^2\Big] + \sum_{\text{dist}_{G_n}(v,u)\le 2r} \mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big].
\end{aligned}$$

Here, we use the fact that if $\text{dist}_{G_n}(u, v) \ge 2r$ then $Z_u(t)$ and $Z_v(t)$ are independent. Therefore,

$$\mathbb{E}\Big[\big(Z_v(t) - \mathbb{E}(Z_v(t))\big)\big(Z_u(t) - \mathbb{E}(Z_u(t))\big)\Big] = 0.$$

To finish the proof note that $0 \le Z_v(t) \le 1$, so an obvious upper bound is $|Z_v(t) - \mathbb{E}(Z_v(t))| \le 1$. Therefore,

$$n^2\text{Var}(S_{n,r}^{(\rho)}(t)) \le n + n^2\varepsilon_{2r}(G_n).$$

Note that all the bounds are independent of t, which finishes the proof. □

Subsequently, we extend this analysis, factoring in the randomness of graph G_n. The following proof emphasizes the stability of local neighborhoods and their role in shaping the variance.

Proof of Lemma 4.4. We first write

$$\text{Var}(S_{n,r}^{(\rho)}(t)) = \mathbb{E}(\text{Var}_\Xi(S_{n,r}^{(\rho)}(t)|G_n)) + \text{Var}(\mathbb{E}_\Xi(S_{n,r}^{(\rho)}(t)|G_n)).$$

The first term can be bounded using Lemma 4.2. For the second term, we use a standard argument we condition over different graph structures of radius r that appear in the r-neighborhood of a uniform random node. Let $\mathcal{H}$ be the set of all graph structures of $B_r(G_n, v_i)$ for $v_i \in V(G_n)$ up to isomorphism. Also for $H^\star \in \mathcal{H}$, recall that $P_r^{(G_n)}(H^\star) = \frac{1}{|V(G_n)|}\sum_{v\in V(G_n)} \mathbb{1}\{B_r(G_n, v) \simeq H^\star\}$ is the probability that the r-neighborhood of a uniform random node in G_n is isomorphic to $H^\star$, and that $p_r^{(n)}(H^\star) = \mathbb{E}[P_r^{(G_n)}(H^\star)]$ is its expectation with respect to the randomness of G_n. Then

$$\mathbb{E}_\Xi(S_{n,r}^{(\rho)}(t)|G_n) = \sum_{H^\star\in\mathcal{H}} P_r^{(G_n)}(H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star). \tag{12}$$

To bound the $\mathcal{H}$, we use the tightness argument. Let $\mathcal{H}_k$ be a subset of $\mathcal{H}$ where each graph has a size of at most k. Then

$$\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n) \le \varepsilon_r(G_n, k) + \sum_{H^\star\in\mathcal{H}_k} P_r^{(G_n)}(H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star). \tag{13}$$

Then using tightness, given any $\delta > 0$ for large enough k and n, $\mathbb{P}(\varepsilon_r(G_n, k) \ge \delta) \le 1-\delta$. Therefore, we can assume that $\text{Var}(\varepsilon_r(G_n, k)) \le 2\delta^2$. Also, using the fact that $S_{n,r}(t) \le 1$, we get $\text{Cov}(\varepsilon_r(G_n, k), \mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n)) \le$

2δ. Therefore,

$$\begin{aligned}
\mathrm{Var}\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n) &\le 2\delta^2 + 2\delta + \mathbb{E}\Big[\Big(\sum_{H^\star:|H^\star|\le k} \mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))\Big)^2\Big] \\
&= 2\delta^2 + 2\delta + \mathbb{E}\Big[\sum_{H^\star:|H^\star|\le k} \mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2\Big] \\
&\quad + 2\sum_{H^\star:|H^\star|\le k}\sum_{H^{\star'}:|H^{\star'}|\le k} \Big(\mathbb{P}_\Xi(t < T^{(r)}(H^\star)|H^\star)\mathbb{P}_\Xi(t < T^{(r)}(H^{\star'})|H^{\star'}) \\
&\qquad \mathbb{E}\Big[(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))(P_r^{(G_n)}(H^{\star'}) - p_r^{(n)}(H^{\star'}))\Big]\Big) \\
&\le 2\delta^2 + 2\delta + \mathbb{E}\Big[\sum_{H^\star}(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2\Big] \\
&\quad + 2\mathbb{E}\Big[\sum_{H^\star:|H^\star|\le k}\sum_{H^{\star'}:|H^{\star'}|\le k} |(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))(P_r^{(G_n)}(H^{\star'}) - p_r^{(n)}(H^{\star'}))|\Big].
\end{aligned}$$

By our stable local neighborhood condition in Definition 3.2, we can bound the last inequality. Let $N_{r,k}$ be the number of rooted graphs of radius r and size k. For a given $\delta' \le \frac{\delta}{N_{r,k}}$, there exists N such that for all $n > N$, and for all $H^\star$, $\mathbb{E}[(P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star))^2] \le \delta'$. Therefore,

$$\mathrm{Var}\Big(\mathbb{E}(S_{n,r}^{(\rho)}(t)|G_n)\Big) \le 3\delta + 4\delta^2. \tag{14}$$

Putting this together with Lemma 4.2, we get

$$\mathrm{Var}(S_{n,r}^{(\rho)}(t)) \le 3\delta + 4\delta^2 + \mathbb{E}[\varepsilon_{2r}(G_n))] + \frac{1}{n}.$$

Finally, the result follows by applying (16) along with the tightness condition to bound $\varepsilon_{2r}(G_n))$. □

A.3 Proof of Theorem 2.1 - Concentration of Epidemic for Deterministic Graphs

For simplicity, we state the proof for the number of susceptible nodes, but all the steps are replicable for infectious and recovered nodes. The first step is to prove that $S_{n,r}^{(\rho)}(t)$ concentrates around the time evolution of epidemic $S_n^{(\rho)}(t)$. In particular, the implications of Lemma 4.2 combined with the Chebyshev inequality shows that for any $\delta' > 0$ and any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S_{n,r}^{(\rho)}(t) - \mathbb{E}_\Xi[S_{n,r}^{(\rho)}(t)]| \ge \delta'\Big) \le \frac{1}{(\delta')^2}\Big(\frac{1}{n} + \varepsilon_{2r}(G_n)\Big).$$

Building upon this, by employing Lemma 4.1 and the Markov inequality, we can further deduce that for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S_{n,r}^{(\rho)}(t) - S_n^{(\rho)}(t)| \ge \delta'\Big) \le \frac{(1-\rho)^r}{\delta'}. \tag{15}$$

Further, Lemma 4.1 also implies that

$$\sup_{t\ge 0}\Big|\mathbb{E}_\Xi[S_{n,r}^{(\rho)}(t)] - \mathbb{E}_\Xi[S_n^{(\rho)}(t)]\Big| \le \mathbb{E}_\Xi\left[\sup_{t\ge 0}\big|S_n^{(\rho)}(t) - S_{n,r}^{(\rho)}(t)\big|\right] \le (1-\rho)^r.$$

Combining the previous three inequalities, we get that for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S_n^{(\rho)}(t) - \mathbb{E}_\Xi[S_n^{(\rho)}(t)]| \ge 2\delta' + (1-\rho)^r\Big) \le \frac{1}{(\delta')^2}\Big(\frac{1}{n} + \varepsilon_{2r}(G_n)\Big) + \frac{(1-\rho)^r}{\delta'}.$$

Now, to finish the proof of the first part, we need to relate the radius discovered by the algorithm to the number of nodes visited. Let $|B_r(G_n, v)|$ be the number of nodes at distance at most r from v. Recall that $n^2\varepsilon_r(G_n)$ is the number of pairs (u, v) such that $\text{dist}_{G_n}(u, v) \le r$. Then

$$\begin{aligned} n^2\varepsilon_r(G_n) &= \sum_v |B_r(G_n, v)| \\ &= \sum_{v:|B_r(G_n,v)|\ge k} |B_r(G_n, v)| + \sum_{v:|B_r(G_n,v)|<k} |B_r(G_n, v)| \\ &\le n^2\varepsilon_r(G_n, k) + kn(1 - \varepsilon_r(G_n, k)). \end{aligned}$$

Here in the last inequality, we use the fact that there are $n\varepsilon_r(G_n, k)$ nodes with $|B_r(G_n, v)| \ge k$. We use the obvious bound of $|B_r(G_n, v)| \le n$ for those, and for the rest of the nodes, we use the bound of $|B_r(G_n, v)| \le k$. Therefore,

$$\varepsilon_r(G_n) \le \varepsilon_r(G_n, k) + \frac{k}{n}. \tag{16}$$

Now it remains to bound the deviation of our estimator with q queries, $\hat{S}^{(\rho)}_{q,r,n}(t)$, with $S^{(\rho)}_{n,r}(t)$. We claim that, for any $t \in [0, \infty]$,

$$\mathbb{P}_\Xi\Big(|S^{(\rho)}_{n,r}(t) - \hat{S}^{(\rho)}_{q,r,n}(t)| \ge \delta\Big) \le 2e^{-2q\delta^2} + \frac{16}{\delta^2}\big(\frac{1}{n} + \varepsilon_{2r}(G_n)\big) \tag{17}$$

Similar as before, define $Z_v(t) = \mathbb{1}\{t < T^{(r)}(G_n, v, \mathcal{M}(G_n))\}$. Let $u_1, u_2, \ldots, u_q$ be the set of initial nodes sampled independently by the algorithm. Then

$$S^{(\rho)}_{n,r}(t) = \frac{1}{n}\sum_{v\in V(G_n)} Z_v(t), \qquad \text{and} \qquad \hat{S}^{(\rho)}_{q,r,n}(t) = \frac{1}{q}\sum_{i=1}^{q} Z_{u_i}(t).$$

We couple the marks drawn in the algorithm with the marks of G_n determining $S^{(\rho)}_{n,r}$. Then

$$\mathbb{P}(Z_{u_i}(t) = 1) = S^{(\rho)}_{n,r}(t),$$

and the sampling of u_i is with replacement. So, $Z_{u_i}(t)$ are independent given $\mathcal{M}(G_n)$. Therefore, using Hoeffding inequality, for $t \in [0, \infty]$

$$\mathbb{P}\Big(|S^{(\rho)}_{n,r}(t) - \hat{S}^{(\rho)}_{q,r,n}(t)| \ge \delta \mid \mathcal{M}(G_n)\Big) \le 2e^{-2q\delta^2}, \tag{18}$$

where the probability is only over the randomness of the starting points of the algorithm u_i. Then, to conclude (17), we can use the variance bound in Lemma 4.2 to decouple the marks of the epidemic with the algorithm. To formalize this, we use the notation $\mathbb{E}_{alg}$ to show expectation over the randomness of the marks drawn by the algorithm, and as before, we use $\mathbb{E}_{G_n}$ for the randomness of G_n and its marks $\mathcal{M}(G_n)$. Also, with the abuse of notation, we use G_n and alg as input of the local approximation $S^{(\rho)}_{n,r}$ and the estimator $\hat{S}^{(\rho)}_{q,r,n}$ to show the corresponding marks. Our goal is to prove the following lower bound,

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t, G_n) - \hat{S}^{(\rho)}_{q,r,n}(t, alg)| \le \delta\Big) \ge 1 - 2e^{-q\delta^2/2} - \frac{16}{\delta^2}(\frac{1}{n} + \varepsilon_{2r}(G_n)). \tag{19}$$

For this purpose, define $E_t = |S^{(\rho)}_{n,r}(t, G_n) - S^{(\rho)}_{n,r}(t, alg)|$ and $\hat{E}_t = |S^{(\rho)}_{n,r}(t, alg) - \hat{S}^{(\rho)}_{q,r,n}(t, alg)|$. Then,

$$\begin{aligned} \mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t, G_n) - \hat{S}^{(\rho)}_{q,r,n}(t, alg)| \le \delta\Big) \ge \\ \mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t, G_n) - \hat{S}^{(\rho)}_{q,r,n}(t, alg)| \le \delta \mid \hat{E}_t \le \delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t \le \delta/2\Big) \ge \\ \mathbb{P}_{\Xi,alg}\Big(E_t + \hat{E}_t \le \delta \mid \hat{E}_t \le \delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t \le \delta/2\Big), \end{aligned}$$

where the last bound is using the triangle inequality. As a result,

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t,G_n)-\hat{S}^{(\rho)}_{q,r,n}(t,alg)|\le\delta\Big)\ge\mathbb{P}_{\Xi,alg}\Big(E_t\le\delta/2\Big)\mathbb{P}_{alg}\Big(\hat{E}_t\le\delta/2\Big).$$

Now, we can apply (18) to bound the coupling between the algorithm and the epidemic, given that the marks are the same.

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t,G_n)-\hat{S}^{(\rho)}_{q,r,n}(t,alg)|\le\delta\Big)\ge\big(1-2e^{-q\delta^2/2}\big)\mathbb{P}_{\Xi,alg}\Big(E_t\le\delta/2\Big).$$

Then we use the variance bound in Lemma 4.2 for the second event:

$$\mathbb{P}_{\Xi,alg}\Big(|S^{(\rho)}_{n,r}(t,G_n)-\hat{S}^{(\rho)}_{q,r,n}(t,alg)|\le\delta\Big)\ge\big(1-2e^{-q\delta^2/2}\big)\big(1-\frac{16}{\delta^2}(\frac{1}{n}+\varepsilon_{2r}(G_n))\big),$$

which proves (17). Combining this with (16), we get the desired result. □

A.4 Concentration of Epidemics for Tight Graphs – Proof of Theorem 4.3.

This is a direct application of Theorem 2.1 along with the definition of the tightness. □

A.5 Examples on Necessity of our Conditions

Two examples are discussed in this section, highlighting the nuanced impact of specific conditions on epidemic estimations. The first illustrates a scenario where the graph does not satisfy the tightness condition and shows how the time evolution of the epidemic does not concentrate in this case. The second example shows where, with a strictly local starting condition, the final size of the epidemic does not concentrate around its mean even if the underlying network is tight.

Example A.1 (Necessity of the tightness condition)**.** The necessity of the condition in Definition 3.1 for local estimation of epidemics becomes apparent when examining specific graph structures. Consider the star graph, wherein a central node connects to $n-1$ peripheral nodes. In this scenario, the final infection size and the time evolution of the epidemic fail to concentrate. To see this, note that except for the ρ fraction of peripheral nodes that are initially infected, the rest of them can only get infected through the central node. Due to its high degree, the central node will, with a high probability, eventually become infected. Assuming the recovery rate to be equal to the transmission rate, the number of nodes to which the central node transmits the disease becomes a uniform random variable within the range of 0 to $(1-\rho)n$. Consequently, without observing the recovery time of the central node, estimating the final infection size or the time evolution becomes infeasible. This example does not satisfy the tightness condition since $|B_2(G_n,v)|=n$ for every node v. Definition 3.1 controls the influence of large-degree nodes and imposes a regularity on the system, leading to more stable and predictable infection spread across the graph. ◀

Example A.2 (Strict locality is not enough for convergence of final size)**.** Consider three distinct graphs, each of size n, where the first is formed by blowing up each node of a 3-regular random graph with a triangle, and the second and third are standard 3-regular random graphs. We add an edge between two random nodes of the first and second graph and two random nodes of the second and third graph. Suppose an initial condition is imposed such that nodes within a triangle are infected while others are susceptible. Under this starting configuration, the first graph becomes entirely infected, while the second and third remain susceptible. The evolution of the epidemic then depends on a single bridging node in the second graph, leading to three potential outcomes for the final infection size: a rapid die-out in the second graph, a linear spread in the second but not the first, or a linear spread in both. Consequently, the final size does not converge to a deterministic value. ◀

B Proof of Theorem 2.3 - Concentration of Epidemic for Random Graphs

The first part of the theorem on the concentration of the epidemic follows the exact same argument as in the proof of Theorem 2.1. To see it, note that Lemma 4.4 is enough to give concentration of $S_{n,r}(t)$. To

deduce the concentration of $S_n(t)$ from it, we note that Lemma 4.1 also applies to random graph models (by conditioning on the drawn graph and using the law of total expectation). To deduce the bound on the local estimator, we can follow the same steps, with the change of applying Lemma 4.4. □

C Convergence of Epidemics on Growing Graphs- Proof Details

C.1 Proof of Lemma 4.6 - Convergence of Local Approximation

The proof is similar to Lemma 4.4 in the sense that we condition on different structures the r ball of a node can take. Recall equation (12), and that P_Ξ is the probability over the graph marks. Also, recall that $s_r^{(\rho)}(t) = \mu_\Xi\Big(\mathbb{1}\{t < T^{(r)}(G,o)\}\Big)$. As before, we can write

$$s_r^{(\rho)}(t) = \sum_{H^\star \in \mathcal{H}} p_r(H^\star)\mathbb{P}_\Xi\big(t < T^{(r)}(H^\star)\big),$$

where $p_r(H^\star) = \mu(\mathbb{1}\{B_r(G,o) \simeq H^\star\})$ is the probability that the r-neighborhood of the limit graph is isomorphic to $H^\star$. Therefore, the left-hand side of the expression of Lemma 4.6 can be written as

$$\sup_{t\geq 0}\left|\mathbb{E}_\Xi[S_{n,r}^{(\rho)}(t)] - s_r^{(\rho)}(t)\right| = \sup_{t\geq 0}\left|\sum_{H^\star\in\mathcal{H}} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\right|. \tag{20}$$

For the rest of the proof, we use tightness and stable local neighborhood criteria for graphs converging locally in probability. These conditions are proved in Appendix C.2. Using the tightness condition, we can choose k large enough such that $\mathbb{P}(\varepsilon_{r,k}(G_n) \leq \delta) \geq 1-\delta$. So, as in Lemma 4.4, if we let $\mathcal{H}_k$ be the set of all locally rooted graphs of size k, then we can approximate the right-hand side of (20) with the sum of $H^\star \in \mathcal{H}_k$. More precisely, using the tightness condition, for any $\delta > 0$, there exists a large enough k, such that

$$\mathbb{P}\Big(\Big|\sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big| - \sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big|\Big| \leq \delta\Big) \geq 1-\delta.$$

By applying this to (20), it is enough to prove the following,

$$\sup_{t\geq 0}\Big|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\Big| \xrightarrow{\mathbb{P}} 0$$

For this purpose, we use the following,

$$\sup_{t\geq 0}\left|\sum_{H^\star\in\mathcal{H}_k} P_\Xi\Big(t < T^{(r)}(H^\star)\Big)\Big(p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big)\right| \leq \sum_{H^\star\in\mathcal{H}_k}\Big|p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big|.$$

So it remains to provide bound on the right-hand side. This is possible by first observing that $|\mathcal{H}_k|$ is a bounded number since there are finitely many graphs of size k. Second, for each $H^\star$, we can use that

$$\Big|p_r^{(G_n)}(H^\star) - p_r(H^\star)\Big| \leq \Big|p_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)\Big| + \Big|p_r^{(n)}(H^\star) - p_r(H^\star)\Big|.$$

The first term goes to zero by stable local neighborhood as proved in Appendix C.3. The second term, $\Big|p_r^{(n)}(H^\star) - p_r(H^\star)\Big| \xrightarrow{\mathbb{P}} 0$ by local convergence in probability (van der Hofstad, 2024, Theorem 2.15 (b)). So, the lemma is proved. □

C.2 Proof of Theorem 2.5 - Convergence of Epidemic

First note that by applying Lemmas 4.1 and 4.4 we get that for any $\delta > 0$, and any r and n large enough, and any $t \in [0, \infty]$,

$$\mathbb{P}\Big(|\mathscr{E}_n(t) - \mathbb{E}_\Xi[\mathscr{E}_{n,r}(t)]| \geq \delta\Big) \leq \delta.$$

Then we can subsequently apply Lemma 4.6 and then Lemma 4.5 to prove convergence of $S_n(t)$ to $s(t)$ uniformly in t. The proof is similar for infectious and recovered nodes. □

C.3 Proof of Theorem 2.6 - Local Approximation of the Limit

The tightness condition follows since the distance of two uniform random nodes increases in convergent graphs. More formally, given a sequence of graphs converging in distribution to a limit, and for any given r, $\lim_{n\to\infty} \varepsilon_r(G_n) = 0$, as demonstrated in (van der Hofstad, 2024, Corollary 2.20).

Also, the stable neighborhood condition Definition 3.2 is satisfied by the criterion of local convergence obtained in (van der Hofstad, 2024, Theorem 2.15 (b)). To formalize this, note that (van der Hofstad, 2024, Theorem 2.15- part b) implies that for any finite rooted graph $H^\star$, and all integers r,

$$P_r^{(G_n)}(H^\star) \xrightarrow{\mathbb{P}} \mu(B_r(G, o) \simeq H^\star).$$

As a result,

$$p_r^{(n)}(H^\star) = \mathbb{E}\big[P_r^{(G_n)}(H^\star)\big] \to \mu(B_r(G, o) \simeq H^\star).$$

Therefore, using a triangle inequality, we get that for any given graph H and integer $r \geq 1$,

$$\mathbb{P}\Big(|P_r^{(G_n)}(H^\star) - p_r^{(n)}(H^\star)| \geq \delta\Big) \leq \mathbb{P}\Big(|P_r^{(G_n)}(H^\star) - \mu(B_r(G, o) \simeq H^\star)| + |\mu(B_r(G, o) \simeq H^\star) - p_r^{(n)}(H^\star)| \geq \delta\Big),$$

where the right side approaches zero when considering the preceding convergences. Therefore, convergent graphs in probability satisfy the stable local neighborhood condition. As a result, we can apply Theorem 2.3, establishing that for given any $\delta > 0$, there exists constants N, q_δ, k_δ such that for any $n > N$,

$$\mathbb{P}\Big(|\hat{\mathscr{E}}^{(\rho)}_{n,q_\delta,k_\delta}(t) - \mathscr{E}_n^{(\rho)}(t)| > \delta\Big) \leq \delta. \tag{21}$$

To finish the proof, it is enough to apply Theorem 2.5, which implies the convergence in probability of $\mathscr{E}_n^{(\rho)}(t)$ to $(s(t), i(t), r(t))$. □

D Proof Details for General Epidemics

D.1 Convergence of General Epidemics – Proof of Corollary 2.7

We follow the proof of Theorem 2.5 step by step and point out what parts of the proofs need to be changed for the general epidemics. After establishing the conclusion of Theorem 2.5, the proof of Theorem 2.6 directly applies in this case.

Recall that the time-varying infectiousness of each node (β_v, τ_v) depends on the ℓ neighborhood of the graph. We start by proving the convergence of the number of susceptible people conditioned on the ℓ neighborhood. As before, we can prove concentration bounds for the number of susceptible nodes and that the truncated epidemics at some r neighborhood give the right bounds.

Our goal is to prove that for any $\delta > 0$ and any $t \in [0, \infty]$,

$$\lim_{r\to\infty} \lim_{n\to\infty} \mathbb{P}(\big|S_n^{(\rho)}(t) - S_{n,r}^{(\rho)}(t)\big| \geq \delta) = 0. \tag{22}$$

As before, we can define $T^{(r)}(v)$ as the time it takes for node v to leave a susceptible state if the epidemic is confined to its r neighborhood. Then, the exact same argument as in the proof of Lemma 4.1, for a uniform random vertex $o_n \in V(G_n)$,

$$\mathbb{P}_\Xi\left(T^{(\infty)}(o_n) \neq T^{(r)}(o_n)\right) \leq (1 - \rho)^r,$$

which implies that

$$\mathbb{E}_{\Xi}\left[\sup_{t\geq 0}|S_n(t)-S_{n,r}(t)|\right]\leq (1-\rho)^r.$$

A similar first-moment bound works for the number of susceptible in the limit.

Now, we can bound $\mathrm{Var}\Big(S_{n,r}(t)\Big)$ as in Lemma 4.4. Note that the bounds we used in the second-moment arguments in the proof of Lemma 4.4 were independent of the specifics of the dynamics of the epidemic, and we only used the fact that $T^{(r)}(v)$ and $T^{(r)}(u)$ of two nodes with $\mathrm{dist}_{G_n}(u,v)>2r$ are independent. Here, this is true if $\mathrm{dist}_{G_n}(u,v)>2r+\ell$, where ℓ is added to ensure the transmission probability densities within all nodes of $B_r(G_n,u)$ and $B_r(G_n,v)$ are independent. The two variance and first-moment bounds prove (22).

Next, by applying the proof steps of Lemma 4.6, we get convergence of $\mathbb{E}_{\Xi}[S_{n,r}(t)]$ to $s_r(t)$, i.e, $\left|\mathbb{E}_{\Xi}[S^{(\rho)}_{n,r}(t)]-s_r(t)\right|\xrightarrow{\mathbb{P}}0$, where the randomness is with respect to G_n. Combining this with the first and second moment results on $S^{(\rho)}_{n,r}(t)$, we get the following convergences for any $t\in[0,\infty]$,

$$S^{(\rho)}_n(t)\xrightarrow{\mathbb{P}}s(t).$$

We now extend our previous proof to calculate the proportion of nodes that reside in state S or any of the states $\mathcal{D}_1,\ldots,\mathcal{D}_i$. Let us define $D^{(i)}=\{S,\mathcal{D}_1,\ldots,\mathcal{D}_i\}$ as the combined set of states encompassing S and $\mathcal{D}_1,\ldots,\mathcal{D}_i$. The term $D^{(i)}_n(t)$ represents the proportion of nodes found in any state within $D^{(i)}$ at a given time t in G_n. Analogously, we define $D^{(i)}(t)$ with respect to the limit graph $(G,o)\sim\mu$. Furthermore, as we previously outlined, $T^{(r)}_i(v,G,\mathcal{M}(G))$ as the time v exits state $\mathcal{D}_i$ and enters $\mathcal{D}_{i+1}$, given that the epidemic is truncated at the r-neighborhood. This is fundamentally equivalent to the shortest route from v to the initial infection state plus the sum $t_1+t_2+\ldots+t_i$ specific to node v (keeping in mind that t_j denotes the transition period from $\mathcal{D}_j$ to $\mathcal{D}_{j+1}$ as determined by β_v). Crucially, considering the given marks, the sole scenario where $T^{(r)}_1(v,G,\mathcal{M}(G))$ differs from $T^{(\infty)}_1(v,G,\mathcal{M}(G))$ is if the shortest route to the initially infected node exceeds a length of r. Otherwise, the contraction time of the disease and $t_1,t_2,\ldots,t_r$ would have already been discerned based on the local neighborhood. Consequently, our preceding proof remains valid in this context: both the variance bound on $\mathbb{1}\{t<T^{(r)}_1\}$ and the convergence to the limit applies to $D^{(i)}(t)$. As a result, $D^{(i)}_n(t)\xrightarrow{\mathbb{P}}D^{(i)}(t)$.

Then the conclusion follows by noting that the number of nodes that in states $\mathcal{D}_i$, can be obtained by the following subtraction $D^{(i)}_n(t)-D^{(i-1)}_n(t)$.

D.2 Proof of Corollary 2.11

Our goal is to prove the convergence of the epidemics with a generalized starting configuration. We show how Lemmas 4.1 and 4.5 can be extended to this case. As before, note that $S_n(t)-S_{n,r}(t)$ only includes nodes that $T^{(\infty)}(o_n)\neq T^{(r)}(o_n)$. For such nodes, there exists a path of length at least $r+1$ from v to an initially infected node.

$$\mathbb{E}_{\Xi}[\sup_{t\geq 0}|\mathscr{E}_n(t)-\mathscr{E}_{n,r}(t)|]=\mathbb{P}_{\Xi}\Big(T^{(\infty)}(o_n)\neq T^{(r)}(o_n)\mid\mathcal{M}(G_n)\Big),$$

where the randomness on the right-hand side is over uniform random node o_n and the infection marks on the initial graph. The right-hand side is essentially the case that a path of length r starting from o_n does not reach I_0. With the local reachability condition in hand, the right-hand side tends to 0 as we increase n and then r.

The second subtle difference in the proof of Theorem 2.6 is that in the variance bound in Lemma 4.4, we have used the fact that $T^{(r)}(u)$ and $T^{(r)}(v)$ are independent if u and v have distance larger than $2r$. With the generalized starting configuration, the independence still holds if u and v have a distance larger than $2r+\ell$. Recall that ℓ is the neighborhood size that initial conditions depend on (through the function P_ℓ). The rest of the proof follows as those of Theorems 2.5 and 2.6 as in other parts of the proof, we do not use the starting configuration.

Table 1: Performance evaluation of the estimator on Copenhagen Dataset. Values in parentheses represent confidence intervals (CI).

Num. of Tests per Query	Absolute Error of Ground Truth vs. Estimated Relative Final Size[a] (CI)	Euclidean Distance of Ground Truth vs. Estimated Time Evolution[b] (CI)	Pearson Correlation of Ground Truth vs. Estimated Time Evolution (CI)
2	0.091 (0.078, 0.104)	0.011 (0.010, 0.012)	0.942 (0.932, 0.952)
3	0.084 (0.071, 0.097)	0.010 (0.010, 0.011)	0.945 (0.937, 0.953)
4	0.086 (0.073, 0.100)	0.010 (0.010, 0.011)	0.952 (0.944, 0.959)
5	0.095 (0.079, 0.111)	0.010 (0.009, 0.010)	0.955 (0.947, 0.962)
6	0.089 (0.074, 0.104)	0.010 (0.010, 0.011)	0.959 (0.953, 0.965)
7	0.106 (0.091, 0.121)	0.010 (0.010, 0.011)	0.959 (0.953, 0.966)
8	0.087 (0.074, 0.099)	0.010 (0.009, 0.011)	0.964 (0.959, 0.969)
9	0.084 (0.071, 0.097)	0.010 (0.009, 0.010)	0.965 (0.959, 0.970)

[a] Defined as $|R_n(\infty) - \hat{R}_{q,k,n}(\infty)|$; the value lies in the range $[0, 1]$.
[b] Defined by the equation $\lim_{T\to\infty} \frac{1}{T} \int_0^T \|\mathscr{E}_n(t) - \hat{\mathscr{E}}_{n,q,k}(t)\|_1$; the value lies in the range $[0, 1]$.

Table 2: Performance evaluation of the estimator on San Francisco Dataset. Values in parentheses represent confidence intervals (CI).

Num. of Tests per Query	Absolute Error of Ground Truth vs. Estimated Relative Final Size (CI)	Euclidean Distance of Ground Truth vs. Estimated Time Evolution (CI)	Pearson Correlation of Ground Truth vs. Estimated Time Evolution (CI)
2	0.103 (0.086, 0.120)	0.058 (0.053, 0.063)	0.595 (0.531, 0.659)
3	0.091 (0.078, 0.105)	0.056 (0.052, 0.060)	0.560 (0.494, 0.625)
4	0.117 (0.099, 0.134)	0.058 (0.053, 0.063)	0.652 (0.598, 0.707)
5	0.107 (0.092, 0.122)	0.060 (0.056, 0.064)	0.542 (0.476, 0.607)
6	0.114 (0.100, 0.127)	0.059 (0.054, 0.065)	0.538 (0.467, 0.610)
7	0.115 (0.098, 0.131)	0.061 (0.056, 0.065)	0.554 (0.494, 0.615)
8	0.105 (0.091, 0.118)	0.062 (0.056, 0.067)	0.510 (0.437, 0.582)
9	0.098 (0.085, 0.111)	0.057 (0.054, 0.061)	0.582 (0.509, 0.654)

Table 3: Performance evaluation of the estimator on Preferential Attachment with 500 nodes. Values in parentheses represent confidence intervals (CI).

Num. of Tests per Query	Absolute Error of Ground Truth vs. Estimated Relative Final Size (CI)	Euclidean Distance of Ground Truth vs. Estimated Time Evolution (CI)	Pearson Correlation of Ground Truth vs. Estimated Time Evolution (CI)
2	0.070 (0.059,0.079)	0.013 (0.013, 0.014)	0.690 (0.657, 0.723)
3	0.063 (0.050,0.074)	0.012 (0.011, 0.013)	0.718 (0.679, 0.757)
4	0.079 (0.065,0.092)	0.012 (0.011, 0.013)	0.718 (0.680, 0.757)
5	0.079 (0.066,0.091)	0.011 (0.011, 0.012)	0.759 (0.720, 0.799)
6	0.073 (0.062,0.083)	0.011 (0.010, 0.012)	0.758 (0.718, 0.797)
7	0.072 (0.061,0.081)	0.011 (0.010, 0.011)	0.776 (0.742, 0.810)
8	0.066 (0.056,0.075)	0.011 (0.011, 0.012)	0.759 (0.729, 0.789)
9	0.067 (0.057,0.074)	0.010 (0.010, 0.011)	0.791 (0.763, 0.818)

Table 4: Performance evaluation of the estimator on Random Geometric Graph with 500 nodes. Values in parentheses represent confidence intervals (CI).

Num. of Tests per Query	Absolute Error of Ground Truth vs. Estimated Relative Final Size (CI)	Euclidean Distance of Ground Truth vs. Estimated Time Evolution (CI)	Pearson Correlation of Ground Truth vs. Estimated Time Evolution (CI)
2	0.080 (0.070, 0.089)	0.063 (0.052, 0.073)	0.855 (0.835, 0.876)
3	0.100 (0.084, 0.11)	0.090 (0.078, 0.102)	0.836 (0.809, 0.864)
4	0.075 (0.063, 0.085)	0.068 (0.058, 0.077)	0.851 (0.830, 0.871)
5	0.075 (0.064, 0.085)	0.077 (0.066, 0.088)	0.864 (0.844, 0.884)
6	0.074 (0.063, 0.084)	0.083 (0.071, 0.095)	0.839 (0.812, 0.865)
7	0.089 (0.078, 0.100)	0.066 (0.056, 0.076)	0.844 (0.820, 0.868)
8	0.076 (0.063, 0.082)	0.074 (0.063, 0.086)	0.863 (0.845, 0.881)
9	0.084 (0.074, 0.092)	0.071 (0.059, 0.084)	0.878 (0.862, 0.894)

Table 5: Absolute error of estimator of the final size of the epidemic the for growing graph size (n). The number of queries is 10, and the testing budget per query is $k = 4$. Confidence intervals are obtained with 1000 simulations.

Graph Size (n)	Preferential Attachment Euclidean Distance (CI)	Random Geometric Graph Euclidean Distance (CI)
500	0.079 (0.066, 0.091)	0.075 (0.064, 0.085)
1000	0.076 (0.064, 0.088)	0.072 (0.061, 0.083)
2000	0.076 (0.063, 0.090)	0.071 (0.061, 0.082)
5000	0.074 (0.064, 0.085)	0.067 (0.059, 0.075)
10000	0.068 (0.058, 0.077)	0.066 (0.058, 0.074)

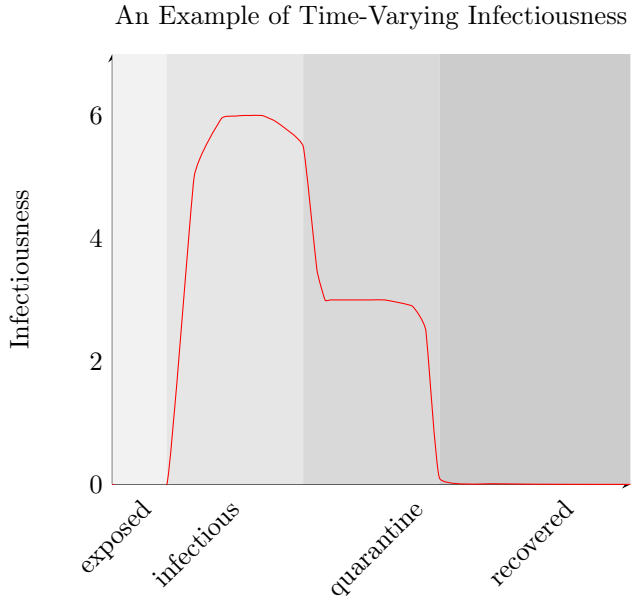

Figure 1: The density of infectiousness over time, illustrating the intervals of exposure, infectiousness, quarantine with reduced transmission rate, and recovery.

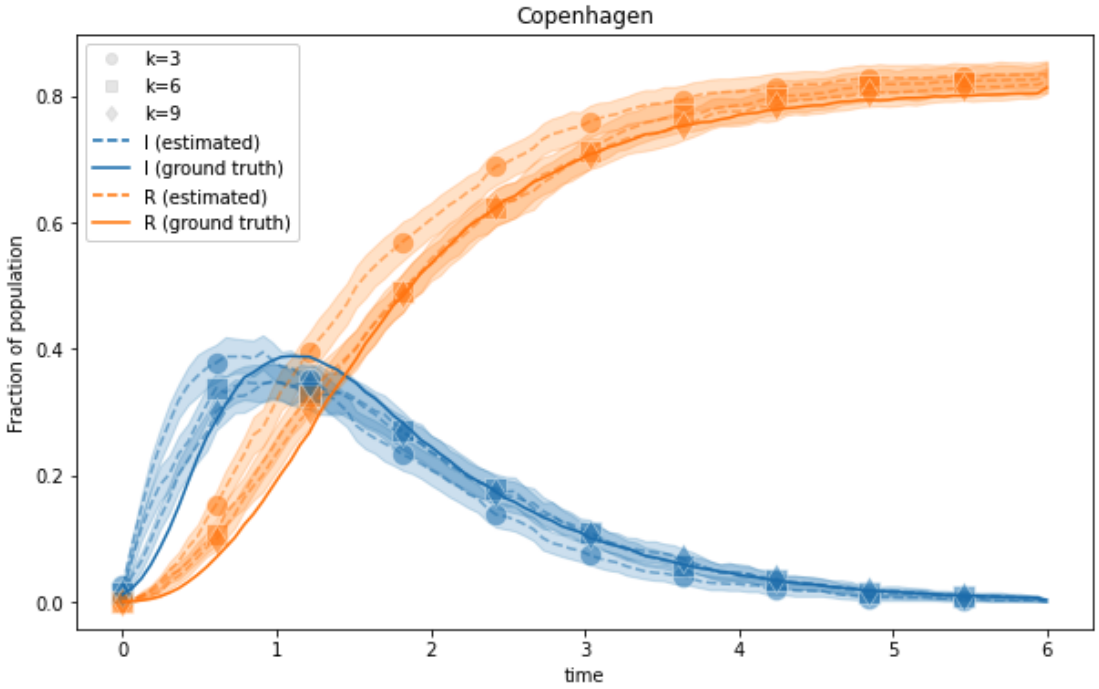

Figure 2: Copenhagen Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

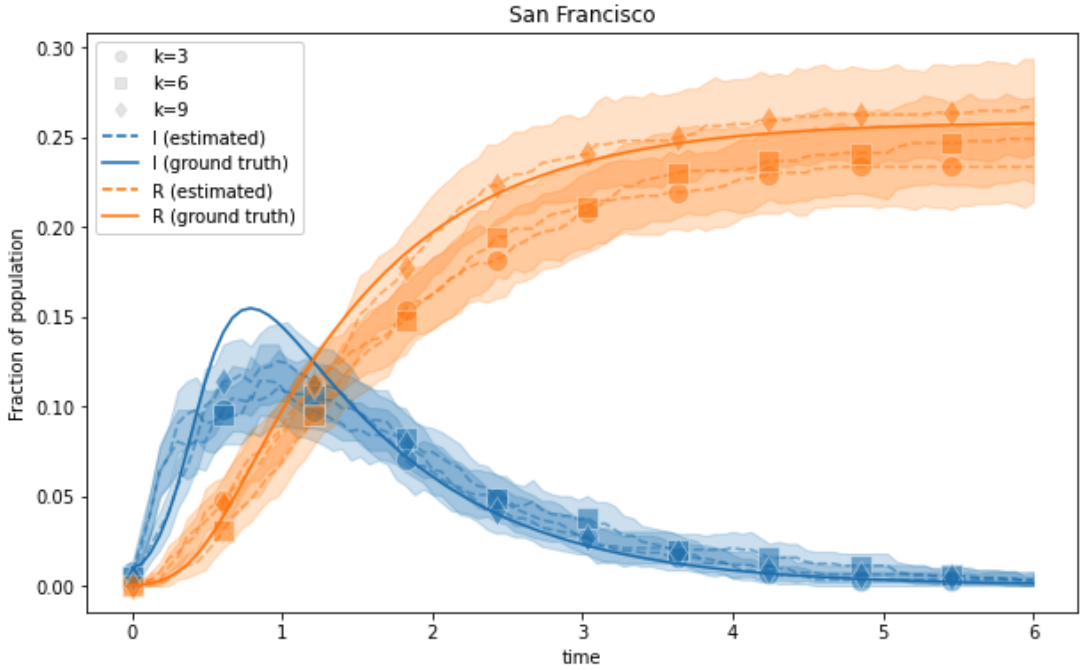

Figure 3: San Francisco Dataset: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

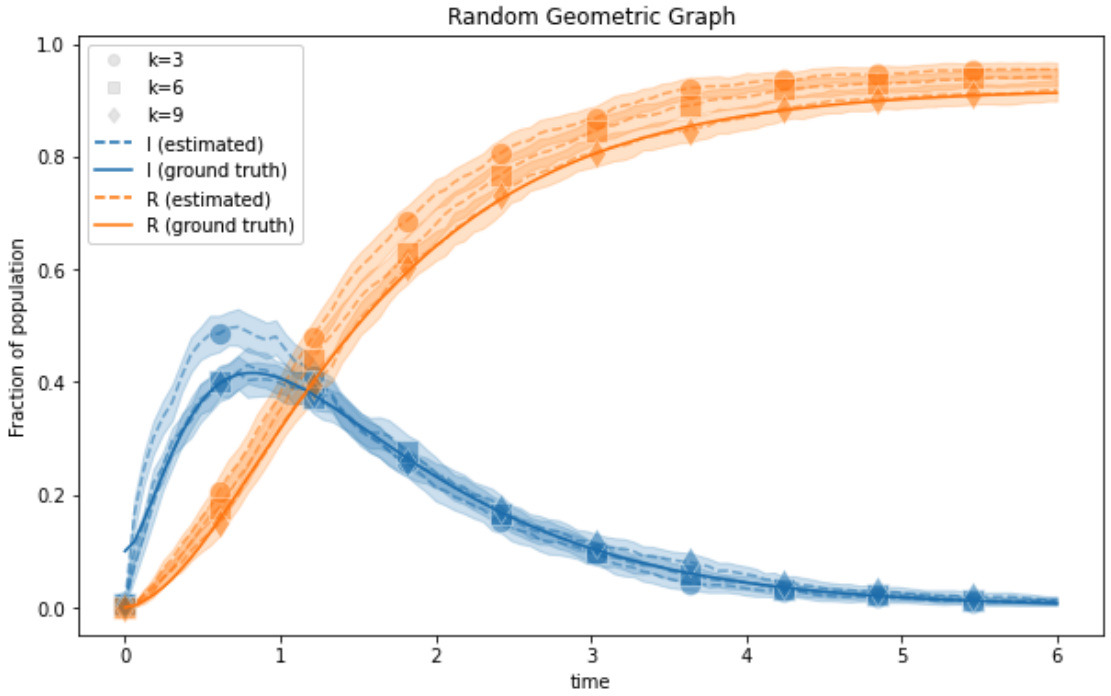

Figure 4: Random Geometric Graph: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

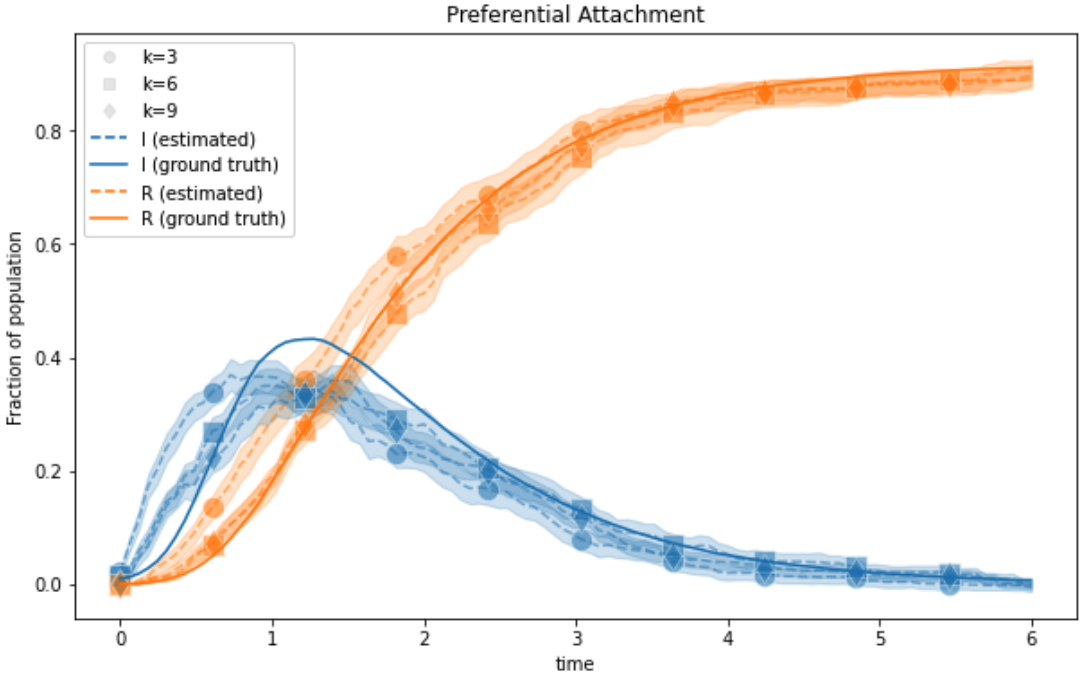

Figure 5: Preferential Attachment: Time Evolution of Epidemic Infections (I) and Recoveries (R) with Estimator Evolution for Various Testing Budgets (k)

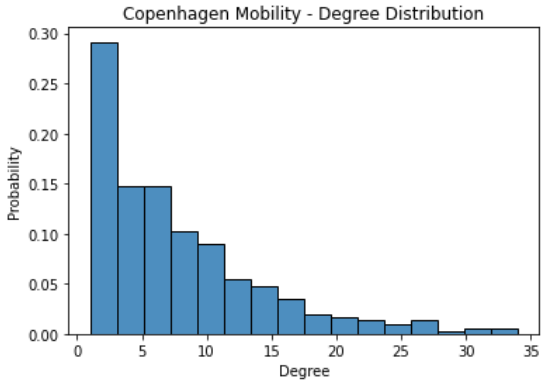

Figure 6: Degree Distribution of Copenhagen Network.

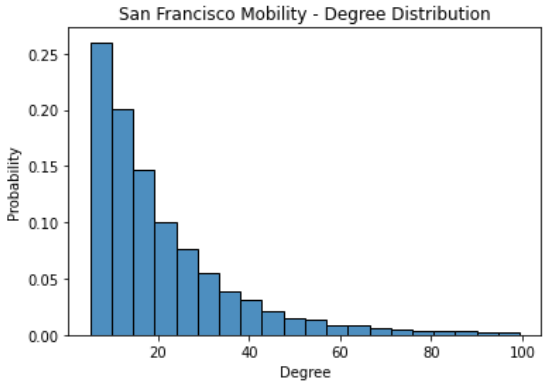

Figure 7: Degree Distribution of San Francisco Mobility Network (Capped at 100 for Depiction).